DCABES 2015

2015 14TH INTERNATIONAL SYMPOSIUM ON DISTRIBUTED COMPUTING AND APPLICATIONS FOR BUSINESS ENGINEERING AND SCIENCE

Published by CPS Conference Publishing Services

CONFERENCE INFORMATION

PAPERS BY SESSION

PAPERS BY AUTHOR

GETTING STARTED

TRADEMARKS

SEARCH

Conference Information

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

---

- ❑ Preface
- ❑ Organizing Committee
- ❑ Reviewers
- ❑ Cover Art (Book version)
- ❑ Title Page (Book version)
- ❑ Copyright Page (Book version)
- ❑ Table of Contents (Book version)
- ❑ Author Index (Book version)
- ❑ Publisher's Information (Book version)

Sessions

- Distributed/Parallel Applications
- Distributed/Parallel Algorithms
- Swarm Intelligence and Applications
- E-Business
- Grid Computing and Cloud Computing
- Network Information Security
- Internet of Things and Applications
- Intelligent Transportation
- Computer Networks and System Architectures
- Big Data Analysis and Decision Support System
- Image Processing
- Technology of Computer Application
- Computational Modeling and Processes
- Soft Computing

Papers by Session

Distributed/Parallel Applications

- Spatial Statistics Parallel Computing Model of Stock
  Zhaojia Dai, Zenli Sun Xinchen, Chuanmei Wang, and Shesheng Zhang
- Weis-Fogh Wave Power Electricity Generation Device Optimization Parallel Model
  Quan Kuang, Xin Chen, and Shesheng Zhang
- The Lager Ship Fluid-Solid Coupling Parallel Algorithm Based on VOSS Mapping Theory
  Yuguang Li, Xin Chen, and Shesheng Zhang
- ZDLC: Towards Automated Software Construction and Migration
  Souheil Khaddaj, Yajna Kumar Makoondlall, Bippin Makoond, Shyam Chivukula, and Kethan Keerthi
- An Optimization Method for Embarrassingly Parallel under MIC Architecture
  Yunchun Li and Xiduo Tian
- Metadata Namespace Management of Distributed File System
  Baoshan Luo, Xinyan Zhang, and Zhipeng Tan
- Research on Distributed Multimedia System in Universities Management Mode
  Linghu Xing-Rong

Papers by Session

- Research on Petri Nets Parallel Algorithm Based on Multi-core PC
  Zhi Zhong, Wenjing Li, Yijuan Su, and Ze-Yu Tang
- GRIB Parallel Design of Civil Aviation Meteorological Data Processing System
  Zhengwei Guo, Yongwei Gao, Yafei Jiang, and Guang Xue

Papers by Session

Distributed/Parallel Algorithms

- A Parallel Algorithm of Green Function with Free Water Surface
  Chao Sun and She-Sheng Zhang
- Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm
  Ding Yanrui, Zhang Zhen, Wang Wenchao, and Cai Yujie
- Performance Analysis for Fast Parallel Recomputing Algorithm under DTA
  Wanfeng Dou and Shoushuai Miao
- Use Pre-record Algorithm to Improve Process Migration Efficiency
  Shan Zhongyuan, Qiao Jianzhong, Lin Shukuan, and Zhang Qiang
- Parallel Algorithm Study of Petri Net Based on Multi-core Clusters
  Wenjing Li, Zhong-Ming Lin, Ying Pan, and Ze-Yu Tang
- Study and Realization on the Partitioning Algorithm of Parallel Subnet of Petri Net System
  Wen-Jing Li, Xiang-Bo Zhang, Yingzhou Bi, and Xuan Wang
- Temporal Logic of Stochastic Actions for Verification of Probabilistic Systems
  Li Jun-tao and Long Shi-gong

Papers by Session

Swarm Intelligence and Applications

- Quantum-Behaved Flower Pollination Algorithm
  Kezhong Lu and Haibo Li
- Multi-objective Flexible Job Shop Schedule Based on Ant Colony Algorithm
  Jiang Xuesong and Tao Qiaoyun
- An Improved QPSO Algorithm Based on Entire Search History
  Ji Zhao, Yi Fu, and Juan Mei
- Proportional Fairness Based Resources Allocation Algorithm for LEO Satellite Networks
  Shuang Xu, Xingwei Wang, and Min Huang
- Quantum-Behaved Particle Swarm Optimization with Cooperative Coevolution for Large Scale Optimization
  Na Tian
- Path Planning for Welding Spot Detection
  Xia Zhu and Renwen Chen
- Characters of a Class of a Rational Difference Equation xn+1=(axnxn-1)/(bxn-1-cxn)
  Xiao Qian, Wang Li-Bin, and Tang Jie

Papers by Session

- A Novel Quantum-Behaved Particle Swarm Optimization Algorithm
  Jing Zhao and Hong Liu
- Solving the Economic Dispatch Problem with Q-Learning Quantum-Behaved Particle Swarm Optimization Method
  Xinyi Sheng and Wnbo Xu
- Structure Learning Algorithm of DBN Based on Particle Swarm Optimization
  Yuansheng Lou, Yuchao Dong, and Huanhuan Ao

Papers by Session

E-Business

- Study of the Influence of Cross-Border Electronic Commerce on Chongqing's Economic Growth
  Qi Wei and Lele Wang
- Collaborative Filtering Recommendation Algorithm Based on MDP Model
  Wang Xingang and Li Chenghao
- Research of O2O-Oriented Service Discovery Method Based on User Context
  She Qiping, Li Qing, Deng Juan, and Wu Zhong
- The Study on the Motivation of T2O E-Commerce Model's Development
  Xu Shuo and Yang Yi
- Effects of RMB Exchange Rate Changes on China's Outward FDI
  Chao Yu
- Mechanism Innovation and Evaluation Model of Wisdom Tourism under the New Situation Based on Survey of Wuxi
  Yawen Cui, Yanping Sun, and Ping Zhu
- Design and Implementation of Logistics Information Management System Based on Web Service
  Hua Jiang, Yuman Li, and Hua Fang

Papers by Session

- Study of Copyright Protection for Merchandise Pictures in E-Commerce
  Liyi Zhang and Chang Liu
- Applying Innovation Resistance Theory to Understand Consumer Resistance of Using Online Travel in Thailand
  Kanjana Jansukpum and Supamas Kettem

Papers by Session

Grid Computing and Cloud Computing

- Cloud Data Migration Method Based on PSO Algorithm
  Geng Yushui and Yuan Jiaheng
- Virtual Machine Migration Strategy in Cloud Computing
  S. Liyanage, S. Khaddaj, and J. Francik
- A VDI System Based on Cloud Stack and Active Directory
  Wei Wei, Yousong Zhang, Yongquan Lu, Pengdong Gao, and Kaihui Mu
- A Fine-Grained and Dynamic MapReduce Task Scheduling Scheme for the Heterogeneous Cloud Environment
  Yingchi Mao, Haishi Zhong, and Longbao Wang

Papers by Session

Network Information Security

- Cryptanalysis of Two Tripartite Authenticated Key Agreement Protocols
  Yang Lu, Quanling Zhang, and Jiguo Li
- Schnorr Ring Signature Scheme with Designated Verifiability
  Xin Lv, Feng Xu, Ping Ping, Xuan Liu, and Huaizhi Su
- Improvement Research Based on Affine Encryption Algorithm
  Yongfeng Wu

Papers by Session

Internet of Things and Applications

- System Design and Obstacle Avoidance Algorithm Research of Vacuum Cleaning Robot
  Li Guangling and Pan Yonghui
- Software Quality Issues and Challenges of Internet of Things
  Jay Kiruthika and S. Khaddaj
- ANN Based High Spatial Resolution Remote Sensing Wetland Classification
  Ke Zun-You, An Ru, and Li Xiang-Juan
- A New Way of Combining RDP and Web Technology for Mobile Virtual Application
  Yousong Zhang, Wei Wei, Pengdong Gao, Yongquan Lu, and Quan Qi
- Kinematics of 3-UPU Parallel Leg Mechanism Used for a Quadruped Walking Robot
  Qifang Gu
- Research and Design of Campus Location Based Service System
  Yugang Hu
- Study on the IOT Architecture and Gateway Technology
  Chang-Le Zhong, Zhen Zhu, and Ren-Gen Huang

Papers by Session

Intelligent Transportation

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System
  Hongxia Wang, Wenkai Guan, Yue Yu, Dongfei Liu, Qiyu Liang, and Yongsheng Yu
- Real-Time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing
  Youwei Yuan, Linliang He, Wanqing Li, Lamei Yan, and M. Mat Deris
- An Application of Fuzzy Rough Sets in Predicting on Urban Traffic Congestion
  Yingchao Shao

Papers by Session

Computer Networks and System Architectures

- Network Performance of EPA Protocol Based on Simulation Tool
  Ping Zhou and Yuqing Feng
- A Distributed Power-Saving Topology Management Scheme in Green Internet
  Jinhong Zhang, Xingwei Wang, and Min Huang
- DMAODV: A MAODV-Based Multipath Routing Algorithm
  Runping Yang and Xia Sun
- Research on Feature Matching of the Field-Based Network Equipment Image
  Juan Fang, Qi Yue, Jianhua Wei, and Junjie Mao
- A Game Based Multi-domain Protection Scheme in WDM Optical Network
  Renzheng Wang, Xingwei Wang, and Min Huang
- A Real-Time Micro-sensor Upper Limb Rehabilitation System for Post-stroke Patients
  Guanhong Tao, Yingfei Sun, Zhipei Huang, and Jiankang Wu
- The Network Model Based on IOCP Memory Control Key Technical Analysis
  Shu Heng
- Bank Partitioning Based Adaptive Page Policy in Multi-core Memory Systems
  Juan Fang, Jiajia Lu, and Min Cai

Papers by Session

- Research and Implementation of Production Rapidly Design and Simulation Verification System Framework
  Fangjing Guan, Haitang Zhu, and Zhifeng Tian
- User Classification Method of P2P Network Based on Clustering
  Shidong Zhang, Yanzhen Li, Zhimin Shao, and Yong Sun
- Design of Ethernet to Optical Fiber Bridge IP Core Based on SOPC
  Yongjun Zhang and Xiangxing Kong

Papers by Session

Big Data Analysis and Decision Support System

- K-Means Clustering Algorithm for Large-Scale Chinese Commodity Information Web Based on Hadoop
  Geng Yushui and Zhang Lishuo
- The Research on Individual Adaptive English Studying of Network Education Platform Based Big Data Technology
  Yanli Song
- A Method for Water Resources Object Identification and Encoding Based on EPC
  Ping Ai, Chuansheng Xiong, Hengli Liao, Dingbo Yuan, and Zhaoxin Yue
- Ensembling Base Classifiers to Improve Predictive Accuracy
  Wen Qingdi
- Expert Achievements Model for Scientific and Technological Based on Association Mining
  Xuexin Qu, Rongjing Hu, Lei Zhou, Liuyang Wang, and Quanyin Zhu
- A Novel Solution of Event Conflict Resolution Based on D-S Evidence Theory
  Xiaojuan Yang and Tao Sun
- A Scene Analysis Model for Water Resources Big Data
  Ping Ai, Zhaoxin Yue, Dingbo Yuan, Hengli Liao, and Chuansheng Xiong

Papers by Session

- An Improved PageRank Algorithm Based on Web Content
  Zhou Hao, Pu Qiumei, Zhang Hong, and Sha Zhihao
- Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification
  Yanxia Yang and Fengli Zhou
- Opinion Leaders Discovering in Social Networks Based on Complex Network and DBSCAN Cluster
  Xiaoli Lin and Wei Han
- Data Analysis of Distributed Application Platform Based on the R Which Apply to Digital Library
  Ningbo Wu and Fan Yang
- A Data Analysis Algorithm of Missing Point Association Rules for Air Target
  Jiang Surong, Lan Jiangqiao, and Yang Yuhai
- A Comprehensive Evaluation System of Association Rules Based on Multi-index
  Shunli Ding, Xin He, and Hong Liang

Papers by Session

Image Processing

- NIB2DPCA-Based Feature Extraction Method for Color Image Recognition
  Zongyue Feng and Jiagang Zhu
- 3D Multi-modality Medical Image Registration Based on Quantum-Behaved Particle Swarm Optimization Algorithm
  Li Hui and Zhu Zhijun
- Research on Medical Image Registration Based on QPSO and Powell Algorithm
  Pan Ting-ting and Zhao Ji
- Face Tracking Algorithm Based on Online Random Forests Detection
  Fang Bao and Yankai Zhang
- Discriminative Sparse Representation and Online Dictionary Learning for Target Tracking
  Huang Yue and Peng Li
- A Secure Blind Watermarking Scheme Based on Embedding Function Matrix
  Wang Xiao and Liu Shuo
- Image Segmentation Method Combines MPM/MAP Algorithm and Geometric Division
  Linghu Yong-Fang and Shu Heng

Papers by Session

- License Plate Location Based on Quantum Particle Swarm Optimization
  Chen Yuping
- A Foreground-Background Segmentation Algorithm for Video Sequences
  Zhou Wei, Peng Li, and Huang Yue
- Processing of Words Labels in Scanned Map Based on Singularity Detection
  Xu Zhipeng and Liu Runqing
- An Improved Live-Wire Freed from the Restriction of the Direct Line Between Seed Points
  Zhou Di and Xu Wenbo
- Quick Capture and Reconstruction for 3D Head
  Chao Lai, Fangzhao Li, and Shiyao Jin
- Locality-Constrained Linear Coding Based on Principal Components of Visual Vocabulary for Visual Object Categorization
  Hongxia Wang, Long Zeng, Dewei Peng, and Feng Geng

Papers by Session

Technology of Computer Application

- Revealing the Structure and Function of P. Pastoris Metabolic Network Using Petri Nets
  Yufang Wang and Dewu Ding
- How to Benchmark Supercomputers
  Gang Xie and Yong-Hao Xiao
- Several Stochastic Gradient Algorithms for Nonlinear Systems with Hard Nonlinearities
  Jia Tang
- Safety Assessment Model Based on Dynamic Bayesian Network
  Yu Feng, Liu Wei, Gao Chunyang, and Tan Lisha
- New Digital Thermostat Development
  Wang Dong and Dai Xunjiang
- Design of Epidemic Monitoring Platform Based on ArcGIS
  Zhang Mei and Yang Zirong
- Improvement of Dynamic Time Warping (DTW) Algorithm
  Yuansheng Lou, Huanhuan Ao, and Yuchao Dong

Papers by Session

Computational Modeling and Processes

- An Optimization Framework Based on Kriging Method with Additive Bridge Function for Variable-Fidelity Problem
  Peng Wang, Yang Li, and Chengshan Li
- Surrogate-Based Optimization for Autonomous Underwater Vehicle's Shell Design
  Huachao Dong, Baowei Song, and Peng Wang
- The Falling Range Prediction Model of Lost Plane
  Kaiyin Zhang, Min Lan, Yan Huang, and Yuguang Li
- Simple Computational Methods for Large Deformation of Plate-Spring End Imposed by Varying Load
  Jun Zhang and Guangyuan Liu
- Large Ship Fluid-Structure Coupling Deformation Calculation Based on Large Deviation Theory
  Xinyun Liu, Xincong zhou, Shesheng Zhang, and Yuguang Li
- Distributed Adaptive Control of Diffusion System Based on Multi-agents
  Tiane Chen, Baotong Cui, and Zaihe Cheng

Papers by Session

- Performance Analysis and Simulation of Vehicle Electronic Stability Control System
  Yang Ying, Liu Weiguo, and Isah Sagir Tukur
- A New Improved Algorithm Based on Three-Stage Inversion Procedure of Forest Height
  Xiang Sun and HongJun Song
- The Research of Feedback-Feedforward Iterative Learning Control in Hydrodynamic Deep Drawing Process
  Songwei Shi
- Electrical Servo Screwdown Control System on Cold Rolling Mill for Traveler Substrate
  Xueyang Yu and Jing Hui
- Research on Tension Control for Coating Line of Optical Films in Dynamic Process
  Chen Ya-Wei and Hui Jing
- A Similarity Model Based on Trend for Time Series
  ShuaiFei Chen, Xin Lv, Lin Yu, YingChi Mao, LongBao Wang, and HongXu Ma
- A Hill-Type Submaximally-Activated Musculotendon Model and Its Simulation
  Lixin Sun, Yingfei Sun, Zhipei Huang, Jiateng Hou, and Jiankang Wu

Papers by Session

- Evaluation of Testing Software Program Based on DEA with Fuzzy Window
  Li Xiang-Juan and Ke Zun-You
- Gender Difference in the Use of Hospitalization Services in Rural China - Evidence from Sichuan Province
  Ye, Shaoxia, and Yin Cong
- The Empirical Analysis on the Influential Factors of Urbanization in Hubei Province Based on the Panel Data
  Yu Qing, Wu Xiaoyuan, Yuan Meng, Xiong Qian, and Jin Shengping

Papers by Session

Soft Computing

- A Risk Probability Model of Study Large Vessel Navigation with Wind and Water Flow
  Li Zhenping, Zhang Shesheng, and Gui Yufeng
- A Research Ship Characteristic Length Model Based on Statistical Theory
  Xuefei Zhang, Xin Chen, Zhenli Sun, Shesheng Zhang, and Yuguang Li
- Chaotic Oscillation Suppression of the Interconnected Power System Based on the Adaptive Back-Stepping Sliding Mode Controller
  Huang Wen-Di, Min Fu-Hong, Wang Zhu-Lin, and Chu Zhou-Jian
- Constructing Kernels for One-Class Support Vector Machine
  Bin Zhang, Jiagang Zhu, and Haobing Tian
- A High Accuracy Spectral Element Method for Solving Eigenvalue Problems
  Weikun Shan and Huiyuan Li
- The Multi-class SVM Is Applied in Transformer Fault Diagnosis
  Liping Qu and Haohan Zhou
- A Modified K-Means Algorithm Based RBF Neural Network and Its Application in Time Series Modelling
  Yiping Jiao, Yu Shen, and Shumin Fei

Papers by Session

- Based on Rough Sets and the Associated Analysis of KNN Text Classification Research
  Guo Aizhang and Yang Tao
- Project Evaluation of Jilin Rural Power Grid Reformation Based on Rough Set and Support Vector Machine
  Du Qiushi, Wang Guan-Nan, and Cong Li
- The Comprehensive Evaluation Index System for Huadian Transformer Substation Address Selection Based on AHP and SVM
  Du Qiushi, Miao Qian, and Cong Li
- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation
  Chenyang Jiang, Feng Xu, Xin Lv, Guoyan Xu, Yingchi Mao, and Longbao Wang
- An Identification Method of News Scientific Intelligence Based on TF-IDF
  Lu Pan, Haibo Tang, Lei Zhou, Liuyang Wang, and Quanyin Zhu
- Estimation of Clusters Number and Initial Centers of K-Means Algorithm Using Watershed Method
  Xiaolong Wang, Yiping Jiao, and Shumin Fei
- The Application of BP Neural Network Algorithm in Optical Fiber Fault Diagnosis
  Yan Shan, Liu Yijuan, and Guan Fangjing

Papers by Session

- Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network
  Jiang Zou, Qingbo Wu, Yusong Tan, Fuhui Wu, and Wenzhu Wang
- Improved Feature Selection Based on Normalized Mutual Information
  Li Yin, Ma Xingfei, Yang Mengxi, Zhao Wei, and Gu Wenqiang
- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic
  Zhang Hong, Sun Chun-Long, Zheng Zi-Jun, An Wei, Zhou De-Qiang, and Wu Jing-Jing

Papers by Author

A

- Ai, Ping
- Aizhang, Guo
- Ao, Huanhuan

B

- Bao, Fang
- Bi, Yingzhou

C

- Cai, Min
- Chen, Renwen
- Chen, ShuaiFei
- Chen, Tiane
- Chen, Xin
- Cheng, Zaihe
- Chenghao, Li
- Chivukula, Shyam
- Chun-Long, Sun
- Chunyang, Gao
- Cong, Yin
- Cui, Baotong
- Cui, Yawen

D

- Dai, Zhaojia
- De-Qiang, Zhou
- Deris, M. Mat
- Di, Zhou
- Ding, Dewu
- Ding, Shunli
- Dong, Huachao
- Dong, Wang

Papers by Author

Papers by Author

J

K

Papers by Author

L

Papers by Author

Papers by Author

- Qiang, Zhang
- Qiaoyun, Tao
- Qing, Li
- Qing, Yu
- Qingdi, Wen
- Qiping, She
- Qiumei, Pu
- Qiushi, Du
- Qu, Liping
- Qu, Xuexin

R

- Ru, An
- Runqing, Liu

S

- Shan, Weikun
- Shan, Yan
- Shao, Yingchao
- Shao, Zhimin
- Shaoxia
- Shen, Yu
- Sheng, Xinyi
- Shengping, Jin
- Shesheng, Zhang
- Shi, Songwei
- Shi-gong, Long
- Shukuan, Lin
- Shuo, Liu
- Shuo, Xu
- Song, Baowei
- Song, HongJun
- Song, Yanli

Papers by Author

Papers by Author

- Wang, LongBao
- Wang, Peng
- Wang, Renzheng
- Wang, Wenzhu
- Wang, Xiaolong
- Wang, Xingwei
- Wang, Xuan
- Wang, Yufang
- Wei, An
- Wei, Jianhua
- Wei, Liu
- Wei, Qi
- Wei, Wei
- Wei, Zhao
- Wei, Zhou
- Weiguo, Liu
- Wenbo, Xu
- Wenchao, Wang
- Wen-Di, Huang
- Wenqiang, Gu
- Wu, Fuhui
- Wu, Jiankang
- Wu, Ningbo
- Wu, Qingbo
- Wu, Yongfeng

X

- Xiang-Juan, Li
- Xiao, Wang
- Xiao, Yong-Hao
- Xiaoyuan, Wu
- Xie, Gang
- Xinchen, Zenli Sun

Papers by Author

- Xingang, Wang
- Xingfei, Ma
- Xing-Rong, Linghu
- Xiong, Chuansheng
- Xu, Feng
- Xu, Guoyan
- Xu, Shuang
- Xu, Wnbo
- Xue, Guang
- Xuesong, Jiang
- Xunjiang, Dai

Y

- Yan, Lamei
- Yang, Fan
- Yang, Runping
- Yang, Xiaojuan
- Yang, Yanxia
- Yanrui, Ding
- Ya-Wei, Chen
- Ye
- Yi, Yang
- Yijuan, Liu
- Yin, Li
- Ying, Yang
- Yong-Fang, Linghu
- Yonghui, Pan
- Yu, Chao
- Yu, Lin
- Yu, Xueyang
- Yu, Yongsheng
- Yu, Yue
- Yuan, Dingbo

Papers by Author

- Yuan, Youwei
- Yue, Huang
- Yue, Qi
- Yue, Zhaoxin
- Yufeng, Gui
- Yuhai, Yang
- Yujie, Cai
- Yuping, Chen
- Yushui, Geng

z

- zhou, Xincong

Z

- Zeng, Long
- Zhang, Bin
- Zhang, Jinhong
- Zhang, Jun
- Zhang, Kaiyin
- Zhang, Liyi
- Zhang, Quanling
- Zhang, Shesheng
- Zhang, She-Sheng
- Zhang, Shidong
- Zhang, Xiang-Bo
- Zhang, Xinyan
- Zhang, Xuefei
- Zhang, Yankai
- Zhang, Yongjun
- Zhang, Yousong
- Zhao, Ji
- Zhao, Jing
- Zhen, Zhang

Papers by Author

Papers by Author

Ai, Ping

- A Method for Water Resources Object Identification and Encoding Based on EPC
- A Scene Analysis Model for Water Resources Big Data

Aizhang, Guo

- Based on Rough Sets and the Associated Analysis of KNN Text Classification Research

Ao, Huanhuan

- Structure Learning Algorithm of DBN Based on Particle Swarm Optimization
- Improvement of Dynamic Time Warping (DTW) Algorithm

Bao, Fang

- Face Tracking Algorithm Based on Online Random Forests Detection

Papers by Author

Bi, Yingzhou

- Study and Realization on the Partitioning Algorithm of Parallel Subnet of Petri Net System

Cai, Min

- Bank Partitioning Based Adaptive Page Policy in Multi-core Memory Systems

Chen, Renwen

- Path Planning for Welding Spot Detection

Chen, ShuaiFei

- A Similarity Model Based on Trend for Time Series

Chen, Tiane

- Distributed Adaptive Control of Diffusion System Based on Multi-agents

Papers by Author

Chen, Xin

- Weis-Fogh Wave Power Electricity Generation Device Optimization Parallel Model
- The Lager Ship Fluid-Solid Coupling Parallel Algorithm Based on VOSS Mapping Theory
- A Research Ship Characteristic Length Model Based on Statistical Theory

Cheng, Zaihe

- Distributed Adaptive Control of Diffusion System Based on Multi-agents

Chenghao, Li

- Collaborative Filtering Recommendation Algorithm Based on MDP Model

Chivukula, Shyam

- ZDLC: Towards Automated Software Construction and Migration

Papers by Author

Chun-Long, Sun

- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic

Chunyang, Gao

- Safety Assessment Model Based on Dynamic Bayesian Network

Cong, Yin

- Gender Difference in the Use of Hospitalization Services in Rural China - Evidence from Sichuan Province

Cui, Baotong

- Distributed Adaptive Control of Diffusion System Based on Multi-agents

Papers by Author

Cui, Yawen

- Mechanism Innovation and Evaluation Model of Wisdom Tourism under the New Situation Based on Survey of Wuxi

Dai, Zhaojia

- Spatial Statistics Parallel Computing Model of Stock

De-Qiang, Zhou

- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic

Deris, M. Mat

- Real-Time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing

Papers by Author

Di, Zhou

- An Improved Live-Wire Freed from the Restriction of the Direct Line Between Seed Points

Ding, Dewu

- Revealing the Structure and Function of P. Pastoris Metabolic Network Using Petri Nets

Ding, Shunli

- A Comprehensive Evaluation System of Association Rules Based on Multi-index

Dong, Huachao

- Surrogate-Based Optimization for Autonomous Underwater Vehicle's Shell Design

Dong, Wang

- New Digital Thermostat Development

Papers by Author

Dong, Yuchao

- Structure Learning Algorithm of DBN Based on Particle Swarm Optimization
- Improvement of Dynamic Time Warping (DTW) Algorithm

Dou, Wanfeng

- Performance Analysis for Fast Parallel Recomputing Algorithm under DTA

Fang, Hua

- Design and Implementation of Logistics Information Management System Based on Web Service

Fang, Juan

- Research on Feature Matching of the Field-Based Network Equipment Image
- Bank Partitioning Based Adaptive Page Policy in Multi-core Memory Systems

Papers by Author

Fangjing, Guan

- The Application of BP Neural Network Algorithm in Optical Fiber Fault Diagnosis

Fei, Shumin

- A Modified K-Means Algorithm Based RBF Neural Network and Its Application in Time Series Modelling
- Estimation of Clusters Number and Initial Centers of K-Means Algorithm Using Watershed Method

Feng, Yu

- Safety Assessment Model Based on Dynamic Bayesian Network

Feng, Yuqing

- Network Performance of EPA Protocol Based on Simulation Tool

Papers by Author

Feng, Zongyue

- NIB2DPCA-Based Feature Extraction Method for Color Image Recognition

Francik, J.

- Virtual Machine Migration Strategy in Cloud Computing

Fu, Yi

- An Improved QPSO Algorithm Based on Entire Search History

Fu-Hong, Min

- Chaotic Oscillation Suppression of the Interconnected Power System Based on the Adaptive Back-Stepping Sliding Mode Controller

Gao, Pengdong

- A VDI System Based on Cloud Stack and Active Directory

Papers by Author

- A New Way of Combining RDP and Web Technology for Mobile Virtual Application

Gao, Yongwei

- GRIB Parallel Design of Civil Aviation Meteorological Data Processing System

Geng, Feng

- Locality-Constrained Linear Coding Based on Principal Components of Visual Vocabulary for Visual Object Categorization

Gu, Qifang

- Kinematics of 3-UPU Parallel Leg Mechanism Used for a Quadruped Walking Robot

Guan, Fangjing

- Research and Implementation of Production Rapidly Design and Simulation Verification System Framework

Papers by Author

Guan, Wenkai

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System

Guangling, Li

- System Design and Obstacle Avoidance Algorithm Research of Vacuum Cleaning Robot

Guan-Nan, Wang

- Project Evaluation of Jilin Rural Power Grid Reformation Based on Rough Set and Support Vector Machine

Guo, Zhengwei

- GRIB Parallel Design of Civil Aviation Meteorological Data Processing System

Papers by Author

Han, Wei

- Opinion Leaders Discovering in Social Networks Based on Complex Network and DBSCAN Cluster

Hao, Zhou

- An Improved PageRank Algorithm Based on Web Content

He, Linliang

- Real-Time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing

He, Xin

- A Comprehensive Evaluation System of Association Rules Based on Multi-index

Heng, Shu

- The Network Model Based on IOCP Memory Control Key Technical Analysis

Papers by Author

- Image Segmentation Method Combines MPM/MAP Algorithm and Geometric Division

Hong, Zhang

- An Improved PageRank Algorithm Based on Web Content
- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic

Hou, Jiateng

- A Hill-Type Submaximally-Activated Musculotendon Model and Its Simulation

Hu, Rongjing

- Expert Achievements Model for Scientific and Technological Based on Association Mining

Hu, Yugang

- Research and Design of Campus Location Based Service System

Papers by Author

Huang, Min

- Proportional Fairness Based Resources Allocation Algorithm for LEO Satellite Networks
- A Distributed Power-Saving Topology Management Scheme in Green Internet
- A Game Based Multi-domain Protection Scheme in WDM Optical Network

Huang, Ren-Gen

- Study on the IOT Architecture and Gateway Technology

Huang, Yan

- The Falling Range Prediction Model of Lost Plane

Huang, Zhipei

- A Real-Time Micro-sensor Upper Limb Rehabilitation System for Post-stroke Patients
- A Hill-Type Submaximally-Activated Musculotendon Model and Its Simulation

Papers by Author

Hui, Jing

- Electrical Servo Screwdown Control System on Cold Rolling Mill for Traveler Substrate

Hui, Li

- 3D Multi-modality Medical Image Registration Based on Quantum-Behaved Particle Swarm Optimization Algorithm

Jansukpum, Kanjana

- Applying Innovation Resistance Theory to Understand Consumer Resistance of Using Online Travel in Thailand

Ji, Zhao

- Research on Medical Image Registration Based on QPSO and Powell Algorithm

Jiaheng, Yuan

- Cloud Data Migration Method Based on PSO Algorithm

Papers by Author

Jiang, Chenyang

- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation

Jiang, Hua

- Design and Implementation of Logistics Information Management System Based on Web Service

Jiang, Yafei

- GRIB Parallel Design of Civil Aviation Meteorological Data Processing System

Jiangqiao, Lan

- A Data Analysis Algorithm of Missing Point Association Rules for Air Target

Jianzhong, Qiao

- Use Pre-record Algorithm to Improve Process Migration Efficiency

Papers by Author

Jiao, Yiping

- A Modified K-Means Algorithm Based RBF Neural Network and Its Application in Time Series Modelling
- Estimation of Clusters Number and Initial Centers of K-Means Algorithm Using Watershed Method

Jie, Tang

- Characters of a Class of a Rational Difference Equation xn+1=(axnxn-1)/(bxn-1-cxn)

Jin, Shiyao

- Quick Capture and Reconstruction for 3D Head

Jing, Hui

- Research on Tension Control for Coating Line of Optical Films in Dynamic Process

Papers by Author

Jing-Jing, Wu

- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic

Juan, Deng

- Research of O2O-Oriented Service Discovery Method Based on User Context

Jun-tao, Li

- Temporal Logic of Stochastic Actions for Verification of Probabilistic Systems

Keerthi, Kethan

- ZDLC: Towards Automated Software Construction and Migration

Kettem, Supamas

- Applying Innovation Resistance Theory to Understand Consumer Resistance of Using Online Travel in Thailand

Papers by Author

Khaddaj, S.

- Virtual Machine Migration Strategy in Cloud Computing
- Software Quality Issues and Challenges of Internet of Things

Khaddaj, Souheil

- ZDLC: Towards Automated Software Construction and Migration

Kiruthika, Jay

- Software Quality Issues and Challenges of Internet of Things

Kong, Xiangxing

- Design of Ethernet to Optical Fiber Bridge IP Core Based on SOPC

Kuang, Quan

- Weis-Fogh Wave Power Electricity Generation Device Optimization Parallel Model

Papers by Author

Lai, Chao

- Quick Capture and Reconstruction for 3D Head

Lan, Min

- The Falling Range Prediction Model of Lost Plane

Li, Chengshan

- An Optimization Framework Based on Kriging Method with Additive Bridge Function for Variable-Fidelity Problem

Li, Cong

- Project Evaluation of Jilin Rural Power Grid Reformation Based on Rough Set and Support Vector Machine
- The Comprehensive Evaluation Index System for Huadian Transformer Substation Address Selection Based on AHP and SVM

Papers by Author

Li, Fangzhao

- Quick Capture and Reconstruction for 3D Head

Li, Haibo

- Quantum-Behaved Flower Pollination Algorithm

Li, Huiyuan

- A High Accuracy Spectral Element Method for Solving Eigenvalue Problems

Li, Jiguo

- Cryptanalysis of Two Tripartite Authenticated Key Agreement Protocols

Li, Peng

- Discriminative Sparse Representation and Online Dictionary Learning for Target Tracking
- A Foreground-Background Segmentation Algorithm for Video Sequences

Papers by Author

Li, Wanqing

- Real-Time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing

Li, Wenjing

- Research on Petri Nets Parallel Algorithm Based on Multi-core PC
- Parallel Algorithm Study of Petri Net Based on Multi-core Clusters

Li, Wen-Jing

- Study and Realization on the Partitioning Algorithm of Parallel Subnet of Petri Net System

Li, Yang

- An Optimization Framework Based on Kriging Method with Additive Bridge Function for Variable-Fidelity Problem

Papers by Author

Li, Yanzhen

- User Classification Method of P2P Network Based on Clustering

Li, Yuguang

- The Lager Ship Fluid-Solid Coupling Parallel Algorithm Based on VOSS Mapping Theory
- The Falling Range Prediction Model of Lost Plane
- Large Ship Fluid-Structure Coupling Deformation Calculation Based on Large Deviation Theory
- A Research Ship Characteristic Length Model Based on Statistical Theory

Li, Yuman

- Design and Implementation of Logistics Information Management System Based on Web Service

Li, Yunchun

- An Optimization Method for Embarrassingly Parallel under MIC Architecture

Papers by Author

Liang, Hong

- A Comprehensive Evaluation System of Association Rules Based on Multi-index

Liang, Qiyu

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System

Liao, Hengli

- A Method for Water Resources Object Identification and Encoding Based on EPC
- A Scene Analysis Model for Water Resources Big Data

Li-Bin, Wang

- Characters of a Class of a Rational Difference Equation xn+1=(axnxn-1)/(bxn-1-cxn)

Papers by Author

Lin, Xiaoli

- Opinion Leaders Discovering in Social Networks Based on Complex Network and DBSCAN Cluster

Lin, Zhong-Ming

- Parallel Algorithm Study of Petri Net Based on Multi-core Clusters

Lisha, Tan

- Safety Assessment Model Based on Dynamic Bayesian Network

Lishuo, Zhang

- K-Means Clustering Algorithm for Large-Scale Chinese Commodity Information Web Based on Hadoop

Liu, Chang

- Study of Copyright Protection for Merchandise Pictures in E-Commerce

Papers by Author

Liu, Dongfei

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System

Liu, Guangyuan

- Simple Computational Methods for Large Deformation of Plate-Spring End Imposed by Varying Load

Liu, Hong

- A Novel Quantum-Behaved Particle Swarm Optimization Algorithm

Liu, Xinyun

- Large Ship Fluid-Structure Coupling Deformation Calculation Based on Large Deviation Theory

Papers by Author

Liu, Xuan

- Schnorr Ring Signature Scheme with Designated Verifiability

Liyanage, S.

- Virtual Machine Migration Strategy in Cloud Computing

Lou, Yuansheng

- Structure Learning Algorithm of DBN Based on Particle Swarm Optimization
- Improvement of Dynamic Time Warping (DTW) Algorithm

Lu, Jiajia

- Bank Partitioning Based Adaptive Page Policy in Multi-core Memory Systems

Lu, Kezhong

- Quantum-Behaved Flower Pollination Algorithm

Papers by Author

Lu, Yang

- Cryptanalysis of Two Tripartite Authenticated Key Agreement Protocols

Lu, Yongquan

- A VDI System Based on Cloud Stack and Active Directory
- A New Way of Combining RDP and Web Technology for Mobile Virtual Application

Luo, Baoshan

- Metadata Namespace Management of Distributed File System

Lv, Xin

- Schnorr Ring Signature Scheme with Designated Verifiability
- A Similarity Model Based on Trend for Time Series
- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation

Papers by Author

Ma, HongXu

- A Similarity Model Based on Trend for Time Series

Makoond, Bippin

- ZDLC: Towards Automated Software Construction and Migration

Makoondlall, Yajna Kumar

- ZDLC: Towards Automated Software Construction and Migration

Mao, Junjie

- Research on Feature Matching of the Field-Based Network Equipment Image

Mao, Yingchi

- A Fine-Grained and Dynamic MapReduce Task Scheduling Scheme for the Heterogeneous Cloud Environment

Papers by Author

- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation

Mao, YingChi

- A Similarity Model Based on Trend for Time Series

Mei, Juan

- An Improved QPSO Algorithm Based on Entire Search History

Mei, Zhang

- Design of Epidemic Monitoring Platform Based on ArcGIS

Meng, Yuan

- The Empirical Analysis on the Influential Factors of Urbanization in Hubei Province Based on the Panel Data

Papers by Author

Mengxi, Yang

- Improved Feature Selection Based on Normalized Mutual Information

Miao, Shoushuai

- Performance Analysis for Fast Parallel Recomputing Algorithm under DTA

Mu, Kaihui

- A VDI System Based on Cloud Stack and Active Directory

Pan, Lu

- An Identification Method of News Scientific Intelligence Based on TF-IDF

Pan, Ying

- Parallel Algorithm Study of Petri Net Based on Multi-core Clusters

Papers by Author

Peng, Dewei

- Locality-Constrained Linear Coding Based on Principal Components of Visual Vocabulary for Visual Object Categorization

Ping, Ping

- Schnorr Ring Signature Scheme with Designated Verifiability

Qi, Quan

- A New Way of Combining RDP and Web Technology for Mobile Virtual Application

Qian, Miao

- The Comprehensive Evaluation Index System for Huadian Transformer Substation Address Selection Based on AHP and SVM

Papers by Author

Qian, Xiao

- Characters of a Class of a Rational Difference Equation xn+1=(axnxn-1)/(bxn-1-cxn)

Qian, Xiong

- The Empirical Analysis on the Influential Factors of Urbanization in Hubei Province Based on the Panel Data

Qiang, Zhang

- Use Pre-record Algorithm to Improve Process Migration Efficiency

Qiaoyun, Tao

- Multi-objective Flexible Job Shop Schedule Based on Ant Colony Algorithm

Qing, Li

- Research of O2O-Oriented Service Discovery Method Based on User Context

Papers by Author

Qing, Yu

- The Empirical Analysis on the Influential Factors of Urbanization in Hubei Province Based on the Panel Data

Qingdi, Wen

- Ensembling Base Classifiers to Improve Predictive Accuracy

Qiping, She

- Research of O2O-Oriented Service Discovery Method Based on User Context

Qiumei, Pu

- An Improved PageRank Algorithm Based on Web Content

Qiushi, Du

- Project Evaluation of Jilin Rural Power Grid Reformation Based on Rough Set and Support Vector Machine

Papers by Author

- The Comprehensive Evaluation Index System for Huadian Transformer Substation Address Selection Based on AHP and SVM

Qu, Liping

- The Multi-class SVM Is Applied in Transformer Fault Diagnosis

Qu, Xuexin

- Expert Achievements Model for Scientific and Technological Based on Association Mining

Ru, An

- ANN Based High Spatial Resolution Remote Sensing Wetland Classification

Runqing, Liu

- Processing of Words Labels in Scanned Map Based on Singularity Detection

Papers by Author

Shan, Weikun

- A High Accuracy Spectral Element Method for Solving Eigenvalue Problems

Shan, Yan

- The Application of BP Neural Network Algorithm in Optical Fiber Fault Diagnosis

Shao, Yingchao

- An Application of Fuzzy Rough Sets in Predicting on Urban Traffic Congestion

Shao, Zhimin

- User Classification Method of P2P Network Based on Clustering

Shaoxia

- Gender Difference in the Use of Hospitalization Services in Rural China - Evidence from Sichuan Province

Papers by Author

Shen, Yu

- A Modified K-Means Algorithm Based RBF Neural Network and Its Application in Time Series Modelling

Sheng, Xinyi

- Solving the Economic Dispatch Problem with Q-Learning Quantum-Behaved Particle Swarm Optimization Method

Shengping, Jin

- The Empirical Analysis on the Influential Factors of Urbanization in Hubei Province Based on the Panel Data

Shesheng, Zhang

- A Risk Probability Model of Study Large Vessel Navigation with Wind and Water Flow

Papers by Author

Shi, Songwei

- The Research of Feedback-Feedforward Iterative Learning Control in Hydrodynamic Deep Drawing Process

Shi-gong, Long

- Temporal Logic of Stochastic Actions for Verification of Probabilistic Systems

Shukuan, Lin

- Use Pre-record Algorithm to Improve Process Migration Efficiency

Shuo, Liu

- A Secure Blind Watermarking Scheme Based on Embedding Function Matrix

Shuo, Xu

- The Study on the Motivation of T2O E-Commerce Model's Development

Papers by Author

Song, Baowei

- Surrogate-Based Optimization for Autonomous Underwater Vehicle's Shell Design

Song, HongJun

- A New Improved Algorithm Based on Three-Stage Inversion Procedure of Forest Height

Song, Yanli

- The Research on Individual Adaptive English Studying of Network Education Platform Based Big Data Technology

Su, Huaizhi

- Schnorr Ring Signature Scheme with Designated Verifiability

Su, Yijuan

- Research on Petri Nets Parallel Algorithm Based on Multi-core PC

Papers by Author

Sun, Chao

- A Parallel Algorithm of Green Function with Free Water Surface

Sun, Lixin

- A Hill-Type Submaximally-Activated Musculotendon Model and Its Simulation

Sun, Tao

- A Novel Solution of Event Conflict Resolution Based on D-S Evidence Theory

Sun, Xia

- DMAODV: A MAODV-Based Multipath Routing Algorithm

Sun, Xiang

- A New Improved Algorithm Based on Three-Stage Inversion Procedure of Forest Height

Papers by Author

Sun, Yanping

- Mechanism Innovation and Evaluation Model of Wisdom Tourism under the New Situation Based on Survey of Wuxi

Sun, Yingfei

- A Real-Time Micro-sensor Upper Limb Rehabilitation System for Post-stroke Patients
- A Hill-Type Submaximally-Activated Musculotendon Model and Its Simulation

Sun, Yong

- User Classification Method of P2P Network Based on Clustering

Sun, Zhenli

- A Research Ship Characteristic Length Model Based on Statistical Theory

Papers by Author

Surong, Jiang

- A Data Analysis Algorithm of Missing Point Association Rules for Air Target

Tan, Yusong

- Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network

Tan, Zhipeng

- Metadata Namespace Management of Distributed File System

Tang, Haibo

- An Identification Method of News Scientific Intelligence Based on TF-IDF

Tang, Jia

- Several Stochastic Gradient Algorithms for Nonlinear Systems with Hard Nonlinearities

Papers by Author

Tang, Ze-Yu

- Research on Petri Nets Parallel Algorithm Based on Multi-core PC
- Parallel Algorithm Study of Petri Net Based on Multi-core Clusters

Tao, Guanhong

- A Real-Time Micro-sensor Upper Limb Rehabilitation System for Post-stroke Patients

Tao, Yang

- Based on Rough Sets and the Associated Analysis of KNN Text Classification Research

Tian, Haobing

- Constructing Kernels for One-Class Support Vector Machine

Papers by Author

Tian, Na

- Quantum-Behaved Particle Swarm Optimization with Cooperative Coevolution for Large Scale Optimization

Tian, Xiduo

- An Optimization Method for Embarrassingly Parallel under MIC Architecture

Tian, Zhifeng

- Research and Implementation of Production Rapidly Design and Simulation Verification System Framework

Ting-ting, Pan

- Research on Medical Image Registration Based on QPSO and Powell Algorithm

Papers by Author

Tukur, Isah Sagir

- Performance Analysis and Simulation of Vehicle Electronic Stability Control System

Wang, Chuanmei

- Spatial Statistics Parallel Computing Model of Stock

Wang, Hongxia

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System
- Locality-Constrained Linear Coding Based on Principal Components of Visual Vocabulary for Visual Object Categorization

Wang, Lele

- Study of the Influence of Cross-Border Electronic Commerce on Chongqing's Economic Growth

Papers by Author

Wang, Liuyang

- Expert Achievements Model for Scientific and Technological Based on Association Mining
- An Identification Method of News Scientific Intelligence Based on TF-IDF

Wang, Longbao

- A Fine-Grained and Dynamic MapReduce Task Scheduling Scheme for the Heterogeneous Cloud Environment
- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation

Wang, LongBao

- A Similarity Model Based on Trend for Time Series

Wang, Peng

- An Optimization Framework Based on Kriging Method with Additive Bridge Function for Variable-Fidelity Problem

Papers by Author

- Surrogate-Based Optimization for Autonomous Underwater Vehicle's Shell Design

Wang, Renzheng

- A Game Based Multi-domain Protection Scheme in WDM Optical Network

Wang, Wenzhu

- Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network

Wang, Xiaolong

- Estimation of Clusters Number and Initial Centers of K-Means Algorithm Using Watershed Method

Wang, Xingwei

- Proportional Fairness Based Resources Allocation Algorithm for LEO Satellite Networks
- A Distributed Power-Saving Topology Management Scheme in Green Internet

Papers by Author

- A Game Based Multi-domain Protection Scheme in WDM Optical Network

Wang, Xuan

- Study and Realization on the Partitioning Algorithm of Parallel Subnet of Petri Net System

Wang, Yufang

- Revealing the Structure and Function of P. Pastoris Metabolic Network Using Petri Nets

Wei, An

- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic

Wei, Jianhua

- Research on Feature Matching of the Field-Based Network Equipment Image

Papers by Author

Wei, Liu

- Safety Assessment Model Based on Dynamic Bayesian Network

Wei, Qi

- Study of the Influence of Cross-Border Electronic Commerce on Chongqing's Economic Growth

Wei, Wei

- A VDI System Based on Cloud Stack and Active Directory
- A New Way of Combining RDP and Web Technology for Mobile Virtual Application

Wei, Zhao

- Improved Feature Selection Based on Normalized Mutual Information

Papers by Author

Wei, Zhou

- A Foreground-Background Segmentation Algorithm for Video Sequences

Weiguo, Liu

- Performance Analysis and Simulation of Vehicle Electronic Stability Control System

Wenbo, Xu

- An Improved Live-Wire Freed from the Restriction of the Direct Line Between Seed Points

Wenchao, Wang

- Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm

Papers by Author

Wen-Di, Huang

- Chaotic Oscillation Suppression of the Interconnected Power System Based on the Adaptive Back-Stepping Sliding Mode Controller

Wenqiang, Gu

- Improved Feature Selection Based on Normalized Mutual Information

Wu, Fuhui

- Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network

Wu, Jiankang

- A Real-Time Micro-sensor Upper Limb Rehabilitation System for Post-stroke Patients
- A Hill-Type Submaximally-Activated Musculotendon Model and Its Simulation

Papers by Author

Wu, Ningbo

- Data Analysis of Distributed Application Platform Based on the R Which Apply to Digital Library

Wu, Qingbo

- Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network

Wu, Yongfeng

- Improvement Research Based on Affine Encryption Algorithm

Xiang-Juan, Li

- ANN Based High Spatial Resolution Remote Sensing Wetland Classification
- Evaluation of Testing Software Program Based on DEA with Fuzzy Window

Papers by Author

Xiao, Wang

- A Secure Blind Watermarking Scheme Based on Embedding Function Matrix

Xiao, Yong-Hao

- How to Benchmark Supercomputers

Xiaoyuan, Wu

- The Empirical Analysis on the Influential Factors of Urbanization in Hubei Province Based on the Panel Data

Xie, Gang

- How to Benchmark Supercomputers

Xinchen, Zenli Sun

- Spatial Statistics Parallel Computing Model of Stock

Papers by Author

Xingang, Wang

- Collaborative Filtering Recommendation Algorithm Based on MDP Model

Xingfei, Ma

- Improved Feature Selection Based on Normalized Mutual Information

Xing-Rong, Linghu

- Research on Distributed Multimedia System in Universities Management Mode

Xiong, Chuansheng

- A Method for Water Resources Object Identification and Encoding Based on EPC
- A Scene Analysis Model for Water Resources Big Data

Xu, Feng

- Schnorr Ring Signature Scheme with Designated Verifiability

Papers by Author

- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation

Xu, Guoyan

- A Novel Changeable Sliding Window Method for Predicting Horizontal Displacement of Dam Foundation

Xu, Shuang

- Proportional Fairness Based Resources Allocation Algorithm for LEO Satellite Networks

Xu, Wnbo

- Solving the Economic Dispatch Problem with Q-Learning Quantum-Behaved Particle Swarm Optimization Method

Xue, Guang

- GRIB Parallel Design of Civil Aviation Meteorological Data Processing System

Papers by Author

Xuesong, Jiang

- Multi-objective Flexible Job Shop Schedule Based on Ant Colony Algorithm

Xunjiang, Dai

- New Digital Thermostat Development

Yan, Lamei

- Real-Time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing

Yang, Fan

- Data Analysis of Distributed Application Platform Based on the R Which Apply to Digital Library

Yang, Runping

- DMAODV: A MAODV-Based Multipath Routing Algorithm

Papers by Author

Yang, Xiaojuan

- A Novel Solution of Event Conflict Resolution Based on D-S Evidence Theory

Yang, Yanxia

- Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification

Yanrui, Ding

- Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm

Ya-Wei, Chen

- Research on Tension Control for Coating Line of Optical Films in Dynamic Process

Papers by Author

Ye

- Gender Difference in the Use of Hospitalization Services in Rural China - Evidence from Sichuan Province

Yi, Yang

- The Study on the Motivation of T2O E-Commerce Model's Development

Yijuan, Liu

- The Application of BP Neural Network Algorithm in Optical Fiber Fault Diagnosis

Yin, Li

- Improved Feature Selection Based on Normalized Mutual Information

Ying, Yang

- Performance Analysis and Simulation of Vehicle Electronic Stability Control System

Papers by Author

Yong-Fang, Linghu

- Image Segmentation Method Combines MPM/MAP Algorithm and Geometric Division

Yonghui, Pan

- System Design and Obstacle Avoidance Algorithm Research of Vacuum Cleaning Robot

Yu, Chao

- Effects of RMB Exchange Rate Changes on China's Outward FDI

Yu, Lin

- A Similarity Model Based on Trend for Time Series

Papers by Author

Yu, Xueyang

- Electrical Servo Screwdown Control System on Cold Rolling Mill for Traveler Substrate

Yu, Yongsheng

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System

Yu, Yue

- Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System

Yuan, Dingbo

- A Method for Water Resources Object Identification and Encoding Based on EPC
- A Scene Analysis Model for Water Resources Big Data

Papers by Author

Yuan, Youwei

- Real-Time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing

Yue, Huang

- Discriminative Sparse Representation and Online Dictionary Learning for Target Tracking
- A Foreground-Background Segmentation Algorithm for Video Sequences

Yue, Qi

- Research on Feature Matching of the Field-Based Network Equipment Image

Yue, Zhaoxin

- A Method for Water Resources Object Identification and Encoding Based on EPC
- A Scene Analysis Model for Water Resources Big Data

Papers by Author

Yufeng, Gui

- A Risk Probability Model of Study Large Vessel Navigation with Wind and Water Flow

Yuhai, Yang

- A Data Analysis Algorithm of Missing Point Association Rules for Air Target

Yujie, Cai

- Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm

Yuping, Chen

- License Plate Location Based on Quantum Particle Swarm Optimization

Yushui, Geng

- Cloud Data Migration Method Based on PSO Algorithm

Papers by Author

- K-Means Clustering Algorithm for Large-Scale Chinese Commodity Information Web Based on Hadoop

Zeng, Long

- Locality-Constrained Linear Coding Based on Principal Components of Visual Vocabulary for Visual Object Categorization

Zhang, Bin

- Constructing Kernels for One-Class Support Vector Machine

Zhang, Jinhong

- A Distributed Power-Saving Topology Management Scheme in Green Internet

Zhang, Jun

- Simple Computational Methods for Large Deformation of Plate-Spring End Imposed by Varying Load

Papers by Author

Zhang, Kaiyin

- The Falling Range Prediction Model of Lost Plane

Zhang, Liyi

- Study of Copyright Protection for Merchandise Pictures in E-Commerce

Zhang, Quanling

- Cryptanalysis of Two Tripartite Authenticated Key Agreement Protocols

Zhang, Shesheng

- Spatial Statistics Parallel Computing Model of Stock
- Weis-Fogh Wave Power Electricity Generation Device Optimization Parallel Model
- The Lager Ship Fluid-Solid Coupling Parallel Algorithm Based on VOSS Mapping Theory
- Large Ship Fluid-Structure Coupling Deformation Calculation Based on Large Deviation Theory

Papers by Author

- A Research Ship Characteristic Length Model Based on Statistical Theory

Zhang, She-Sheng

- A Parallel Algorithm of Green Function with Free Water Surface

Zhang, Shidong

- User Classification Method of P2P Network Based on Clustering

Zhang, Xiang-Bo

- Study and Realization on the Partitioning Algorithm of Parallel Subnet of Petri Net System

Zhang, Xinyan

- Metadata Namespace Management of Distributed File System

Zhang, Xuefei

- A Research Ship Characteristic Length Model Based on Statistical Theory

Papers by Author

Zhang, Yankai

- Face Tracking Algorithm Based on Online Random Forests Detection

Zhang, Yongjun

- Design of Ethernet to Optical Fiber Bridge IP Core Based on SOPC

Zhang, Yousong

- A VDI System Based on Cloud Stack and Active Directory
- A New Way of Combining RDP and Web Technology for Mobile Virtual Application

Zhao, Ji

- An Improved QPSO Algorithm Based on Entire Search History

Zhao, Jing

- A Novel Quantum-Behaved Particle Swarm Optimization Algorithm

Papers by Author

Zhen, Zhang

- Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm

Zhenping, Li

- A Risk Probability Model of Study Large Vessel Navigation with Wind and Water Flow

Zhihao, Sha

- An Improved PageRank Algorithm Based on Web Content

Zhijun, Zhu

- 3D Multi-modality Medical Image Registration Based on Quantum-Behaved Particle Swarm Optimization Algorithm

Zhipeng, Xu

- Processing of Words Labels in Scanned Map Based on Singularity Detection

Papers by Author

Zhong, Chang-Le

- Study on the IOT Architecture and Gateway Technology

Zhong, Haishi

- A Fine-Grained and Dynamic MapReduce Task Scheduling Scheme for the Heterogeneous Cloud Environment

Zhong, Wu

- Research of O2O-Oriented Service Discovery Method Based on User Context

Zhong, Zhi

- Research on Petri Nets Parallel Algorithm Based on Multi-core PC

Zhongyuan, Shan

- Use Pre-record Algorithm to Improve Process Migration Efficiency

Papers by Author

Zhou, Fengli

- Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification

Zhou, Haohan

- The Multi-class SVM Is Applied in Transformer Fault Diagnosis

Zhou, Lei

- Expert Achievements Model for Scientific and Technological Based on Association Mining
- An Identification Method of News Scientific Intelligence Based on TF-IDF

Zhou, Ping

- Network Performance of EPA Protocol Based on Simulation Tool

Papers by Author

zhou, Xincong

- Large Ship Fluid-Structure Coupling Deformation Calculation Based on Large Deviation Theory

Zhou-Jian, Chu

- Chaotic Oscillation Suppression of the Interconnected Power System Based on the Adaptive Back-Stepping Sliding Mode Controller

Zhu, Haitang

- Research and Implementation of Production Rapidly Design and Simulation Verification System Framework

Zhu, Jiagang

- NIB2DPCA-Based Feature Extraction Method for Color Image Recognition
- Constructing Kernels for One-Class Support Vector Machine

Papers by Author

Zhu, Ping

- Mechanism Innovation and Evaluation Model of Wisdom Tourism under the New Situation Based on Survey of Wuxi

Zhu, Quanyin

- Expert Achievements Model for Scientific and Technological Based on Association Mining
- An Identification Method of News Scientific Intelligence Based on TF-IDF

Zhu, Xia

- Path Planning for Welding Spot Detection

Zhu, Zhen

- Study on the IOT Architecture and Gateway Technology

Papers by Author

Zhu-Lin, Wang

- Chaotic Oscillation Suppression of the Interconnected Power System Based on the Adaptive Back-Stepping Sliding Mode Controller

Zi-Jun, Zheng

- A Modified Dynamic Window Approach to Obstacle Avoidance Combined with Fuzzy Logic

Zirong, Yang

- Design of Epidemic Monitoring Platform Based on ArcGIS

Zou, Jiang

- Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network

Zun-You, Ke

- ANN Based High Spatial Resolution Remote Sensing Wetland Classification

Papers by Author

- Evaluation of Testing Software Program Based on DEA with Fuzzy Window

Papers by Author

A B C D E F G

H I J K L M N

O P Q R S T U

V W X Y Z

Prev Home

IEEE Computer Society Conference Publications Operations Committee

IEEE Computer Society Publications

The world-renowned IEEE Computer Society publishes, promotes, and distributes a wide variety of authoritative computer science and engineering texts. These books are available from most retail outlets. Visit the CS Store at *http://www.computer.org/portal/site/store/index.jsp* for a list of products.

IEEE Computer Society *Conference Publishing Services* (CPS)

The IEEE Computer Society produces conference publications for more than 250 acclaimed international conferences each year in a variety of formats, including books, CD-ROMs, USB Drives, and on-line publications. For information about the IEEE Computer Society's *Conference Publishing Services* (CPS), please e-mail: cps@computer.org or telephone +1-714-821-8380. Fax +1-714-761-1784. Additional information about *Conference Publishing Services* (CPS) can be accessed from our web site at: *http://www.computer.org/cps*

IEEE Computer Society / Wiley Partnership

The IEEE Computer Society and Wiley partnership allows the CS Press *Authored Book* program to produce a number of exciting new titles in areas of computer science and engineering with a special focus on software engineering. IEEE Computer Society members continue to receive a 15% discount on these titles when purchased through Wiley or at: *http://wiley.com/ieeecs*. To submit questions about the program or send proposals, please e-mail jwilson@computer.org or telephone +1-714-816-2112. Additional information regarding the Computer Society's authored book program can also be accessed from our web site at:
http://www.computer.org/portal/pages/ieeecs/publications/books/about.html

Revised: 21 January 2008

CPS Online is our innovative online collaborative conference publishing system designed to speed the delivery of price quotations and provide conferences with real-time access to all of a project's publication materials during production, including the final papers. The ***CPS Online*** workspace gives a conference the opportunity to upload files through any Web browser, check status and scheduling on their project, make changes to the Table of Contents and Front Matter, approve editorial changes and proofs, and communicate with their CPS editor through discussion forums, chat tools, commenting tools and e-mail.

The following is the URL link to the ***CPS Online*** Publishing Inquiry Form:
http://www.ieeeconfpublishing.org/cpir/inquiry/cps_inquiry.html

Trademarks

ORIENTATION

This Electronic Guide file contains hypertext links to individual article files. Links are represented by colored text (e.g., a name or title); clicking on the text activates the link. Before you start browsing and using the information on this CD-ROM, you will need to install Adobe Acrobat Reader + Search 7.0. If you already have Acrobat Reader installed on your system, make sure it is version 6.0 or higher and includes the Search plug-in. The README.TXT file found on the root directory of this CD-ROM provides additional information. In many instances, we refer to the "menu bar" and "tool bar", shown here for reference.

Menu Bar:

Tool Bars:

Be sure to read the following on how to achieve the best performance with this electronic guide.

RECOMMENDATIONS FOR OPTIMAL PERFORMANCE

In order to take full advantage of the performance capabilities of this collection, we recommend that you do the following.

To make navigation and searching easier, we strongly recommend changing the following Acrobat Search

Preferences (found under Edit > Preferences > General > Search on the menu bar.) In the dialog box shown for Acrobat Search Preferences, change the following: A. Turn ON "Always use advanced search options" so that additional subject fields are visible when specifying search criteria. If for some reason this preference option is not present on your system, check to see that you have the search plug-in installed.

The Search icon: 

will be present on the Acrobat tool bar if the function is properly installed. Specifics of the Search function are described later in this section. B. Change "Maximum number of documents returned in Results" from "100" to "1000" in the field provided.) This allows the maximum number of hits to be displayed during a search.

2. Articles may have text outside normal print-area defaults. We recommend selecting "Fit to paper" in the print menu (File > Print) to capture the complete image for your printout. These settings will become your new default.

RETURNING TO THE ELECTRONIC GUIDE

After viewing an article, there are three methods of returning to the electronic guide.

* **IMPORTANT NOTE**: To keep documents from closing after linking to a PDF file of a paper, go to Edit Menu > Preferences > General. There should be no checkmark next to "open cross-document links in same window."

a) Select Document > Go Back Doc from the menu bar.

b) Click (as needed) on the Previous View button on the tool bar.

1. Select View > Go To > Previous View from the menu bar.

2. Click (as needed) on the Previous View button on the tool bar.

3. Select from the open file list from the "Window" menu.

SELECTING TEXT AND GRAPHICS

To select text or graphics, the appropriate select tool must be visible and selected. The select tools share the same space on the tool bar. The text select tool is visible by default. To select the graphics select tool, you must press and hold on the text select tool. A pop-up will display the select tools, which you can then select. See the Adobe Acrobat Reader 7.0 Guide (Help > Reader Guide) for more information on these tools.

NAVIGATION BUTTONS

Section Map

The current section is shown at the top of each page. The "path" to this section is shown at the right. Clicking these text buttons moves you to the start of that section.

Next Page

Click to advance to the next page in the section.

Previous Page

Click to go back to the previous page in the section. (The Page Up and Page Down keys perform the same functions as the Next and Previous Page buttons.)

Fast Forward Pages

Click to advance (jump) multiple pages in the section.

Fast Back Pages

Click to go back (jump) multiple pages in the section.

ACROBAT TOOLBAR

Navigation

Go to Last Page
Go Forward One Page
Go Backward One Page
Go to First Page

History

Go to Next View (One Link)
Go to Previous View (One Link)

ACROBAT TOOLBAR (continued)

Zoom

Zoom Out
Zoom In
Move Page Image

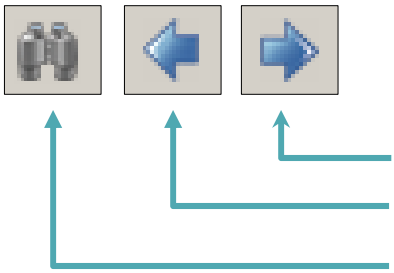
Selection

Select Image
Select Table
Select Text

ACROBAT TOOLBAR (continued)

Page View

Fit Width
Fit Page
Actual Size

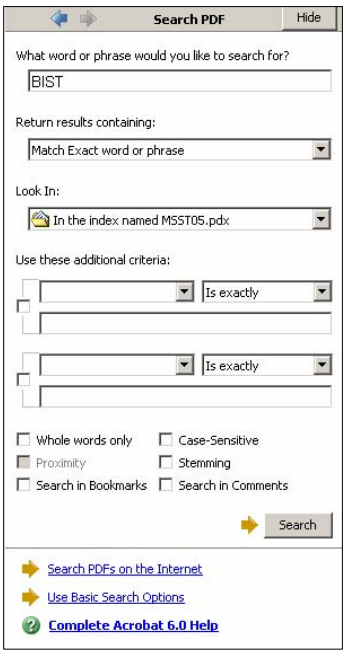


PERFORMING A "SEARCH"

Choosing the "Search" tool bar button or Search menu item (Edit > Search) opens a window on the right side of the screen. Search scans linearly through the currently open Acrobat file from the cursor forward. If the Electronic Guide PDF is open, Search will scan the entire electronic Guide for a match to your text. Type a text string in the field provided, check the appropriate options and press the "Search" button. Reader then highlights the first instance of the text string. To look at the next "hit", click on the link in the Results window or Results menu item (Edit > Search Results > Next Result from Menu).

The "Using Advanced Search Options" link at the bottom of the search window (toggles with Basic Options) provides access to the more powerful full-text search engine (if you installed Acrobat Reader from this CD-ROM). Choose "Look In:" to specify the search location. Selecting the "index.pdx" file, typing a term in the text box at the top of the Search dialog box, and pressing the "Search" button causes a full-text search of all words in the body of papers in the collection. The Advanced Search Options window is shown on the next page.

Choosing "Return results containing" or "Using these additional criteria" searches for hits in only those fields. If you do not find files that should appear in the results list, Acrobat might not be attached to the correct index file. To check, go to "Look In:" > "Select Index" for a list of available indexes. If this collection is not listed, press the "Add..." button and look in the root directory of the CD-ROM for a file called "index.pdx". Click on that file to add it to the list. See the Acrobat Reader Guide (on Help menu) for more complete instructions on selecting appropriate options, constructing boolean queries, etc.



DCABES 2015

2015 14TH INTERNATIONAL SYMPOSIUM ON DISTRIBUTED COMPUTING AND APPLICATIONS FOR BUSINESS ENGINEERING AND SCIENCE

Published by CPS Conference Publishing Services

CONFERENCE INFORMATION

PAPERS BY SESSION

PAPERS BY AUTHOR

GETTING STARTED

TRADEMARKS

SEARCH

PROCEEDINGS

14th International Symposium on Distributed Computing and Applications for Business, Engineering and Science

—— DCABES 2015 ——

PROCEEDINGS

14th International Symposium on Distributed Computing and Applications for Business, Engineering and Science

—— DCABES 2015 ——

18–24 August 2015
Guiyang, China

IEEE Computer Society Order Number: E5535
BMS Part Number: CFP1520K-USB
ISBN: 978-1-4673-6593-2

Additional copies may be ordered from:

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

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

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

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

Editorial production by Randall Bilof
Cover art production by Mark Bartosik

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

DCABES 2015 Preface

The Distributed Computing and Applications for Business, Engineering and Science (DCABES) conference series has become an important international event for the presentation of research on various applications and related computing environments of distributed and grid computing. The 14th International Symposium on Distributed Computing and Applications for Business, Engineering and Science (DCABES 2015) will be held at Jiangnan University and Guizhou University of Finance and Economics. DCABES conferences have also been held in Hangzhou, the Three Gorges Dam Project region (Yichang), Greenwich (UK), Wuhan, Hong Kong, Kingston (UK), and Xianning.

The conference topics include not only traditional topics such as parallel and distributed computing, but also intelligent computing and other topics. We were pleased that the DCABES 2015 conference received a great number of submissions covering a wide range of topics, such as Parallel/Distributed Computing, Image Processing, Network Technology and Information Security, E-Commerce and E-Business, Information Processing, Internet of Things, Swarm Intelligence, Soft Computing, etc. All papers contained in this proceedings have been peer reviewed and carefully chosen by members of the Program Committee, Proceedings Editorial Board, and external reviewers. Papers acceptance or rejection was based on the majority opinions of the referees. All papers contained in this proceedings give us a glimpse of what future technology and applications are being researched in the distributed computing area of the world.

We would like to thank all members of the Scientific Committee, the Local Organizing Committee, the Proceedings Editorial Board, and external reviewers for selecting the papers. It was indeed a pleasure to work with them and obtain their suggestions. We are also grateful to Professor Horacio González-Vélez, Professor Yuhui Shi, Dr. Vasile Palade, and Professor Xiao-Jun Wu for contributing keynote speeches to the conference.

Sincere thanks are also given to the China Ministry of Education (MOE), National Nature Science Foundation of China (NSFC), Jiangnan University, Guizhou University of Finance and Economics, and the Easychair website.

Finally, we also want to thank Dr. Wei Fang, Dr. Jun Sun, Dr. Yanrui Ding, Dr. Zhiguo Chen, Mr. Xinhong Song, and Miss Yiqiong Yuan for their efforts in conference organizing activities. Without their time and effort, this conference could not have been organized smoothly.

Prof. Wenbo Xu, Jiangnan University, China
DCABES 2015 Chair

Prof. Jianzhong Chen, Guizhou University of Finance and Economics, China
Prof. Xiao-Jun Wu, Jiangnan University, China
DCABES 2015 Co-Chairs

DCABES 2015 Organizing Committee

Conference Chair

Prof. Wenbo Xu, Jiangnan University, China

Conference Co-Chairs

Prof. Jianzhong Chen, Vice President, Guizhou University of Finance and Economics
Prof. Xiao-Jun Wu, Jiangnan University, China

Steering Committee

Prof. Craig C. Douglas, University of Wyoming Mathematics Department, Yale University Computer Science Department
Prof. Qing-Ping Guo, Wuhan University of Technology, Wuhan, China (Co-chair)
Prof. Choi-Hong Lai, University of Greenwich, UK (Co-chair)
Thomas Tsui, Chinese University of Hong Kong, Hong Kong, China
Prof. Wenbo Xu, Jiangnan University, Wuxi, China

Program Committee

Prof. Craig C. Douglas, Mathematics Dept., Univ. of Wyoming; Computer Science Dept., Yale Univ., USA
Prof. Jianzhong Chen, Vice President, Guizhou University of Finance and Economics, China
Prof. Zirong Yang, Guizhou University of Finance and Economics, China
Prof. C.-H. Lai, School of Computing and Mathematical Sciences, University of Greenwich, UK
Prof. Q.P. Guo, Wuhan Univ. of Technology, School of Computer Science and Technology, Wuhan, China
Prof. Frederic Magoules, Applied Mathematics and Systems, Ecole Centrale Paris, France
Prof. Albert Y. Zomaya, Chair Professor in Centre for Distributed and High Performance Computing, School of Information Technologies, The University of Sydney, Australia
Professor Dr. Hai Jin, Dean of School of Computer Science and Technology, HUST, Wuhan, China
Prof. Maurício Vieira Kritz, Department of Applied and Computational Mathematics, National Laboratory for Scientific Computation, Petropolis RJ, Brasil
Prof. Peter Jimack, Pro Dean for Research, Faculty of Engineering, Dean of the Faculty of Engineering, Professor of Scientific Computing, Director of the Computational PDEs Unit, University of Leeds, UK
Prof. W.B. Xu, School of Internet of Things, Jiangnan University, Wuxi, China
Prof. Xiao-Chuan Cai, Department of Computer Science, College of Engineering and Applied Science, University of Colorado at Boulder, USA
Prof. Jianwen CAO, Laboratory of Parallel Computing, Institute of Software Chinese Academy of Sciences, Beijing, China
Prof. Director, Chi XueBing, Supercomputing Center, Computer Network Information Center, Chinese Academy of Sciences, China
Prof. Yakup Paker, Department of Computer Science, Queen Mary University of London, London, UK
Dr. Turgay Altilar, Istanbul Technical University, Istanbul, Turkey

Dr. Prof. Souheil Khaddaj, Faculty of Computing, Information Systems and Mathematics, Kingston University, UK

Dr. Andrew A. Chien, Director and SAIC Chair Professor, Center for Networked Systems, University of California, San Diego, USA

Dr. Pui-Tak Ho, Deputy Director of Computer Centre, The University of Hong Kong, HK

Prof. John W. T. Lee, Department of Computing, The Hong Kong Polytechnic University, HK

Prof. Lishan Kang, Department of Computer Science and Technology, University of Geosciences, China

Prof. David Keyes, Applied Physics & Applied Mathematics, Columbia University, USA; Dean, Mathematical and Computer Sciences and Engineering Named Professor, Applied Mathematics and Computational Science, King Abdullah University of Science and Technology

Dr. Chen, Professor Wei, School of Automation, Wuhan University of Technology, Wuhan, China

Dr. Xiao, Professor Xinping, School of Science, Wuhan University of Technology, China

Prof. Wenjing Li, Dept. of Computers, Guangxi Normal University, Nanning, Guangxi, China

Prof. Shesheng Zhang, School of Science, Wuhan University of Technology, China

Ping Lin, Professor, Department of Mathematics, University of Dundee, UK

Prof. Michael K. Ng, Department of Mathematics, The University of Hong Kong, HK

Prof. Sun Jiachang, Parallel Software Research and Development Center, Institute of Software, Academy of Science, China

Dr. Alfred Loo, Associate Professor, Department of Computing and Decision Science, Lingnan Univ., HK

Dr. Man Leung Wong, Department of Computing and Decision Sciences, Lingnan University, HK

Dr. Prof. Peter Kacsuk, the Laboratory of the Parallel and Distributed Systems in the Computer and Automation Research Institute, Hungarian Academy of Sciences, Hungary

Prof. Stefan Vandewalle, Computer Science Department, Katholieke Universiteit Leuven, Belgium

Dr. Robert Lovas, MTA SZTAKI–Computer and Automation Research Institute, Hungarian Academy of Sciences, Hungary

Dr. Associate Professor Faouzi Alaya Cheikh, Department of Computer Science and Media Technology, Gjovik University College, Norway

Dr. Prof. Nikos Christakis, Department of Applied Mathematics, University of Crete, Heraklion, Greece

Prof. Haixin Lin, Faculty of Information Technology and Systems., Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands

Prof. Anne Trefethen, E-Research Centre, University of Oxford, UK

Prof. Rassul Ayani, School of Information and Communication Technologies (ICT), Royal Institute of Technology (KTH), Sweden

Prof. Zhihui Du, Department of Computer Science and Technology, Tsinghua University, China

Prof. Meiqing Wang, College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China

Prof. Yuhua Liu, Professor/Tutor, Dept. of Computer Science, Central China Normal Univ., Wuhan, China

Rongcong Xu, College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China

Prof. Youwei Yuan, School of Computer & Software, Hangzhou Dianzi University, Hangzhou, China

Prof. V. P. Kutepov, Moscow Power Engineering Institute, Russia

Prof. Yi Pan, Department of Computer Science, Georgia State University, USA

Prof. Alan Davies, School of Physics, Astronomy and Mathematics, University of Hertfordshire, UK

Prof. Peter Sloot, Section Computational Science, University of Amsterdam, The Netherlands
Prof. Franck Cappello, CNRS, Universite Paris-Sud, France
Prof. Simon Cox, Computational Methods in the Computational Engineering Design Research Group, School of Engineering Sciences, UK
Prof. Laurie Cuthbert, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
Prof. Alex Shafarenko, Dept. Comp. Science, University of Hertfordshire, UK
Prof. Liu Dan, Computer Science and Technology Department, Henan Mechanic and Electric Engineering College, China
Prof. Xiaojun Tong, Dept. of Computers, Harbin Institute of Technology, Wei Hai, China
Dr. Liu, Professor Dan, China Criminal Police University, China
Dr. Ray Hyatt, Jr., Brown University, USA
Dr. Lamine M. Aouad, School of Computer Science & Informatics, University College Dublin, Ireland
Dr. Thi-Mai-Huong Nguyen, Applied Mathematics and Systems Laboratory, Ecole Centrale Paris, France
Dr. Haiwu He, School of Computer and Information, Hohai University, China
Dr. Alan J. Davies, Mathematics Department, University of Hertfordshire, UK
Prof. Ziyue Tang, Air Force Radar Institute, Wuhan, China
Dr. Prof. Yalchin Efendiev, Institute for Science Computation, College of Science, Texas A&M Univ., USA
Dr. Prof. Yuhui Shi, Dept. of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China
Dr. Prof. Qifeng Yang, Dept. of Electric Commerce, School of Economy, Wuhan University of Technology, Wuhan, China
Dr. Prof. Dongwoo Sheen, Dept. of Mathematics, Seoul National University, Korea
Professor Rule Hjelsvold, Faculty of Technology, Gjovik University College, Norway
Dr. Morten Irgens, Dean of Department of Computer Science and Media Technology, Gjovik University College, Norway
Dr. Prof. Mohamed Kamel, Canada Research Chair in Cooperative Intelligent Systems, Pattern Analysis and Machine Intelligence, Electrical and Computer Engineering, University of Waterloo, Canada
Dr. Prof. Dexin Zhan, Dean of Hydrodynamics Institute, Wuhan University of Technology, Wuhan, China
Dr. Prof. Liyi Zhang, Dean of Department of Electronic Commerce, School of Information Management, Wuhan University, Wuhan, China
Pedro Leite da Silva Dias, Director of National Laboratory for Scientific Computing, Ministry of Science and Technology, Petropolis, RJ, Brazil
Dr. Professor Shengwu Xiong, School of Computer Science & Technology, Wuhan University of Technology China
Dr. Professor Xinming Tan, School of Computer Science & Technology, Wuhan Univ. of Technology, China
Dr. Professor Shu Gao, School of Computer Science & Technology, Wuhan Univ. of Technology, China
Dr. Professor Hongxing Liu, School of Computer Science & Technology, Wuhan Univ. of Technology, China
"Neal" Naixue Xiong, PhD, Department of Computer Science, Georgia State University
Lecturer Yucheng Guo, School of Computer Science & Technology, Wuhan Univ. of Technology, China
Prof. Shitong Wang, School of Digital Media, Jiangnan University, China

Prof. Xiao-Jun Wu, School of Internet of Things Engineering, Jiangnan University, China
Prof. Zhen-Qiu Ning, Jiangnan University Wuxi, China
Prof. Yi-Hua Zhu, Zhejiang University of Technology, Hangzhou, China

Local Organizing Committee

Prof. Wenbo Xu, Jiangnan University, Wuxi, China (Chair)
Prof. Guo-Dong Shi, Changzhou University, China
Prof. Fei Liu, Jiangnan University, Wuxi, China
Prof. Yuan Liu, Jiangnan University, Wuxi, China
Prof. Shi-Tong Wang, Jiangnan University, Wuxi, China
Prof. Xiao-Jun Wu, Jiangnan University, Wuxi, China
Prof. Hong-Yuan Wang, Changzhou University, China
Prof. Xi-Huang Zhang, Jiangnan University, Wuxi, China
Prof. Hong-Wei Ge, Jiangnan University, Wuxi, China
Prof. Zhen-Qiu Ning, Jiangnan University, Wuxi, China
Prof. Yi-Hua Zhu, Zhejiang University of Technology, Hangzhou, China

Secretariat

Mr. Xinhong Song, Jiangnan University, Wuxi, China
Miss. Yiqiong Yuan, Jiangnan University, Wuxi, China

DCABES 2015 Reviewers

Chaofeng Li, Jiangnan University
Cheng-Yuan Li, Jiangnan University
Choi-Hong Lai, University of Greenwich
Craig Douglas, University of Wyoming, Yale University
Dan Liu, Criminal Police University
Di Zhou, Jiangnan University
Guo Qingping, Wuhan University of Technology
Guodong Shi, Changzhou University
Hong-Wei Ge, Jiangnan University
Xinming Tan, Wuhan University of Technology China
Jun Sun, Jiangnan University
Man Leung Wong, Lingnan University.HK
Meiqing Wang, Fuzhou University
Ning Zhenqiu, Jiangnan University
Shu Gao, Wuhan University of Technology China
Shitong Wang, Jiangnan University
Souheil Khaddaj, Kingston University
Hongyuan Wang, Changzhou University
Jianwen Cao, Research and Development Centre for Parallel Algorithms and Software
Xiaojun Wu, Jiangnan University
Yi-Hua Zhu, Zhejiang Universitry of Technology
Zhiguo Chen, Jiangnan University
Wenbo Xu, Jiangnan University
Jianzhong Chen, Guizhou University of Finance and Economics
Frederic Magoules, Ecole Centrale Paris
Shengwu Xiong, Wuhan University of Technology China
Hongxing Liu, Wuhan University of Technology China
Yuhui Shi, Xi'an Jiaotong-Liverpool University

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

DCABES 2015

Table of Contents

Distributed/Parallel Applications

Distributed/Parallel Algorithms

Swarm Intelligence and Applications

E-Business

Grid Computing and Cloud Computing

Network Information Security

Internet of Things and Applications

Intelligent Transportation

Computer Networks and System Architectures

Big Data Analysis and Decision Support System

Image Processing

Technology of Computer Application

Computational Modeling and Processes

Soft Computing

Spatial Statistics Parallel Computing Model of Stock

Zhaojia Dai, Zenli Sun XinChen
Wuhan University of Technology, Wuhan, 430070,China
e-mail: ChenXin@qq.com

Chuanmei Wang*, Shesheng Zhang
Wuhan University of Technology, Wuhan, 430070
e-mail:wchuanmei@163.com

***Abstract*— The research of stare is useful for country and people. A space share rising statistical parallel model is built by using statistical theory. The formula of correlation between stock prices is given. The regression equation with space coordination is obtained..**

Keywords-ship, stock, statistics, space coordination, regression.

I. INTRODUCTION

In 2014, the stock trading volume is over 74 trillion Yuan in china stock markets. It is also a hug trading volume in the world. And in china, trading volume will be increase after 2014. The government, stock company, and science researcher are interested in stock market. The statistical method was often used to find the way forecast stock price, such as regression model [1], increment ratio of rational function model [2], supervised classification of share price trends model [3], and time series prediction based on pattern classification algorithm [4]. The nonlinear forecast method is also found to predict share price [5]. In theory, the nonlinear method is better than linear one. But it is not easy to predict by using simple nonlinear model. In order to reduce prediction error, more complex model is built. Such as recurrent neural network and a hybrid model [6], complex efficient system for stock market prediction [7]. As we know, complex model often need long CPU time of calculating, some times may be one week or long. People can't wait such too long time to get results. One way to short wall time is using parallel computer.

We will consider space share rising statistical parallel model to find some rule of researching stock market. In the section II, the basic analysis stock data is given. In the section III, a space statistical model is discussed. In the section IV, the way of parallel calculation is shown. In the section V, the calculation results are obtain and shows in the table or figure drawn.

II. STOCK DATA COLLECTED

Suppose x is north latitude, y is east longitude, r is height measured from see lever, and Q(t,x,y,r) is stock price random process with time t and space coordination (x,y,r). The observation data Qj= Q(tj,xj,yj,rj), j=1,2,..,N, is come from stock market. For example, the stock of E-Wushang(000501) has north latitude x=30.35o, east longitude y=114.17o, and height r=27m, the stock price is shown in Fig.2.1. From figure, we find, the stock price is waved with time, some time is go up, some time go down..

Table.2.1 shows stock ID and coordination. There are 39 stock companies chosen from China. Some of them is located in north of China, some of them is in south. The data is collected time from 2008 -2-20 to 2015-2-17, almost 7 years. The linear interpolation method is used for the data missed, or stock was not transacted. The stock price rise percent pj is defined as:

$$p_j = 100\frac{Q_j - Q_{j-1}}{Q_j} \qquad (2.1)$$

The stock price rise percent of E-Wuishang(000501) is shown in Fig.2.2. from figure, we find, the percent value almost random. Such random data will consider at next section.

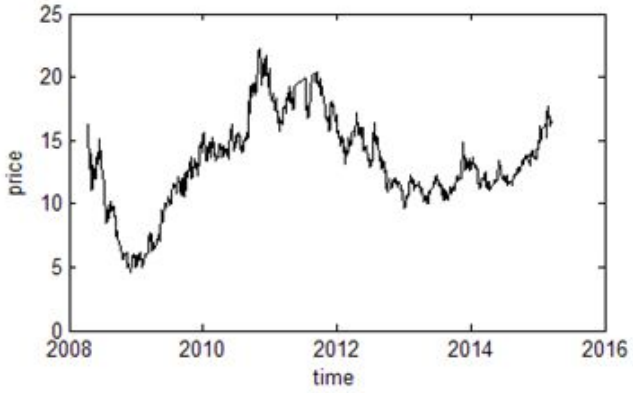


Fig.2.1 E-Wuishang(000501) stock price varied with time

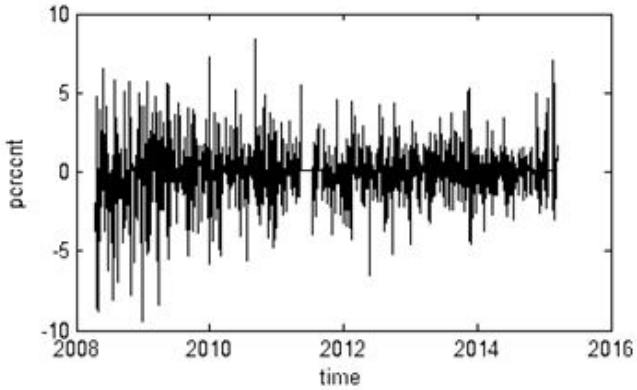


Fig.2.2 The stock rise percent varied with time

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.8

III. SPACE STATISTICAL MODEL

.Let P represented stock price rising percent. x1=t is time(unit is year), x2=x is north latitude, x3=y is east longitude, x4=r is height measured from see lever, and xj,j=5,6,7,..,M are other economic statistical data. Let

$$p = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_M x_M + \varepsilon \qquad (3.1)$$

Here is normal error distribution with mean zero. By using data (pkj,x1s,...,xMs), s=1,2,...,N. We have：

$$p_s = a_0 + a_1 x_{1_s} + a_2 x_{2_s} + \cdots + a_M x_{M_s} + \varepsilon_s \qquad (3.2)$$

s=1,2,,,.N

Let

$$X_k = \begin{pmatrix} 1 & x_{11} & \cdots & x_{M1} \\ 1 & x_{12} & \cdots & x_{M2} \\ \cdots & \cdots & \ddots & \cdots \\ 1 & x_{1N} & \cdots & x_{MN} \end{pmatrix} \qquad A_k = \begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_M \end{pmatrix}$$

$$P = \begin{pmatrix} p_1 \\ p_2 \\ \cdots \\ p_N \end{pmatrix} \qquad \Omega_k = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_N \end{pmatrix} \qquad (3.3)$$

We have matrix formula：

$$P = XA + \Omega \qquad (3.4)$$

It has solution of minimum error

$$A = (X^T X)^{-1} X^T P \qquad (3.5)$$

The sum of absolute error is

$$||\Omega|| = |\varepsilon_1| + ... + |\varepsilon_N| \qquad (3.6)$$

We choose l stock companies, the correlation matrix between l stocks is

$$\mathrm{R} = \begin{pmatrix} R_{11} & R_{12} & ... & R_{1l} \\ R_{21} & R_{22} & ... & R_{2l} \\ ... & ... & ... & ... \\ R_{l1} & R_{l2} & ... & R_{ll} \end{pmatrix} \qquad (3.7)$$

where

$$\mathrm{R}_{ij} = \frac{\mathrm{E}(P_i P_j) - \mathrm{E}(P_i)\mathrm{E}(P_j)}{\sqrt{\mathrm{D}(P_i)\mathrm{D}(P_j)}} \qquad (3.8)$$

And Pj is stock rise percent of jth stock company, j=1,2,..,l.

IV. PARALLEL CALCULATE

Data come from stock market is huge. For example, there are over 5000 stock companies in China, and stock market is open over 200 days for a year, 10 year is open 2000 days. Every day open 4 hours, every hour have 3600 s. every second at lest two data. So that over 10 year, we have the number of data 5000*2000*4*3600*2=2.88*1011. Based on such huge data, we must calculate correlative matrix, average value, variance, and other statistics data. The work will spend long CPU time. The wall time will be short if we use parallel computer. The steps of parallel calculating are:

(1)divide stock company to M parts, every part is calculated by one computer;

(2)Calculate matrix X and matrix R in every part, and then exchange data between computers.

(3)Calculate invert matrix buy using parallel algorithm.

(4)Calculate regression coefficients and error.

(5)Output results, and check it.

M=39 processors are chosen in the calculation. The results are obtained from above steps. Fig.4.1 shows stock average of stock rise percent for 39 stocks. We find most of them is below zero. That means, during 2008 to 2015, the stock price is decrease for most time. Only few of them are go up. Fig.4.2 shows stock varies for 39 stock companies. The value of variance is from 1.5 to 4.0. That means, the stock price is wave too much during this time. The average and variance are also calculated., We find, the larger variance has larger abstract average value. Table 4.1 shows correlative value for five stocks. We find the correlative value is close 0.4. Fig.4.3 shows correlation of 20th stock varies with other stock. The conclusion is same, which is the correlative value is close 0.4.

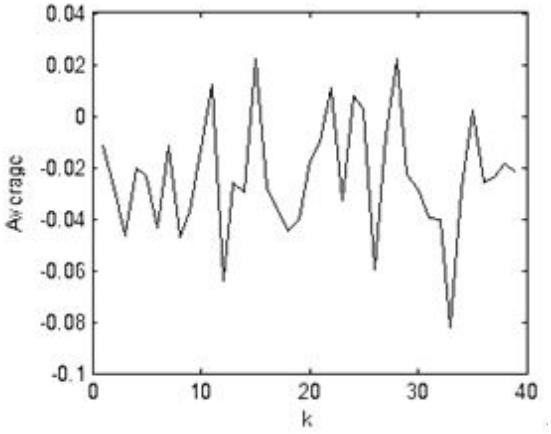


Fig.4.2 Stock average varies with subscribe number k

V. RESULTS AND DISCUSS

We know x represents north latitude, y represents east longitude, t represents time. Let stock rise percent has below formula:

$$p(t) = a_0(t) + a_x(t)x + a_y(t)y + a_r(t)r + \varepsilon \qquad (5.1)$$

That is, the coefficients of coordination are the function of time t. By using data collected, the coefficients of x varied with time t is shown in Fig.5.1. From the figure, we find, the coefficients of x are less than zero, and the average value is -0.0063172. That means, the stock in the north of china will be rise, if stock in the south is fall. Otherwise, the stock in

the north of china will be fall under condition of the stock in the south is rise. The coefficient of y is shown in the Fig.5.2. From figure, we find, the coefficient is larger than zero, the average value is 0.011017. That means, the stock in the north of china will be rise, if stock in the south is also rise.

Let

$$p = a_0 + a_x x + a_y y + a_t t + \varepsilon \qquad (5.2)$$

By using regression method, we can obtained linear regression equation with x,y,t varies:

$$p = -42.338 - 0.0013499x + 0.0010147y + 0.020999t \qquad (5.3)$$

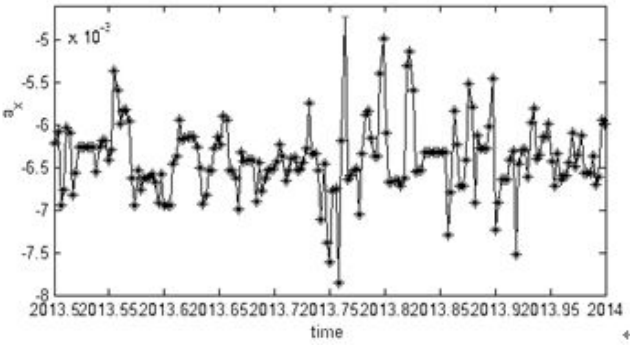


Fig.5.1 Coefficient of x varied with time, average value is -0.0063172

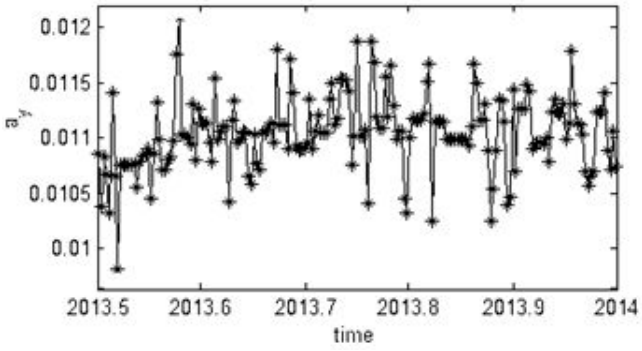


Fig.5.2 Coefficient of y varied with time, average value is 0.011017.

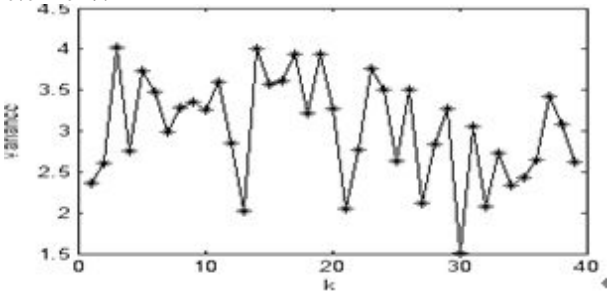
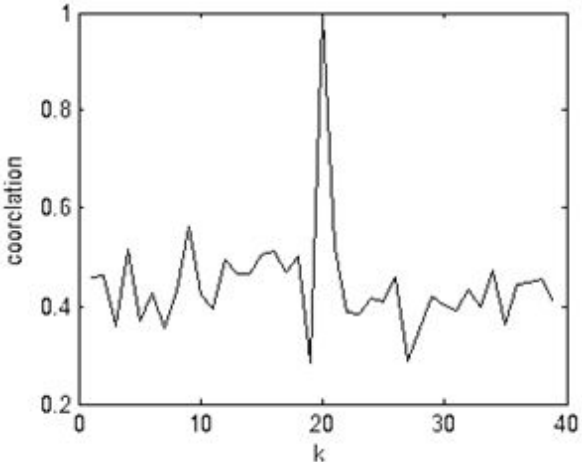


Fig.4.2 Stock varies with subscribe number k



VI. CONCLUSIONS

This paper considers stock distribution with space coordination (north latitude, east longitude, height measured from see lever) and time. The rising percent function is defined with stock price varied with time. The average and variance is calculated for 39 stocks. The regression formulas are obtained both for time and for space coordination.

ACKNOWLEDGMENT

This work was financially supported by the Fundamental Research Funds for the Central Universities (WUT: 2014- la-035) and Humanity and Social Science foundation for Youth of Ministry of Education of China (14YJCZH143, 14YJCZH173).

REFERENCES

[1] ..Shesheng Zhang, mathematical model of stock internal market value[J], Application of statistics and management , 1995(05), 24-27.

[2] Shesheng Zhang,A mathematical model for increment ratio of rational function and its application in share prediction[J]. Jour. of Wuhan transportation university, 1995(01), 101-105.

[3] Zhanggui Zeng, Hong Yan. Supervised classification of share price trends[J]. Information Science,2008(178):2943-2956.

[4] Z.Zeng, M.N. Fu, H. Yan, Time series prediction based on pattern classification[J]. Artificial Intelligence in Engineering,15(2001):61–69.

[5] Hai Zhu, Haiyan Wang, Nonlinear forecast model for China stock market[J], Journal of Anhui University of Technology and Science, 2004(02), 10-13.

[6] Rather, Akhter, Recurrent neural network and a hybrid model for prediction of stock returns[J],

[7] Expert Systems with Applications, 2015,v 42, n 6, p 3234-3241

[8] Hussein, Ashraf S., et al. An efficient system for stock market prediction[J], Advances in Intelligent Systems and Computing, 2015, v 323, p 871-882

Table 2.1 Stock ID and coordination

j	stock	x	y	r	j	Stock	x	y	r
1	(000501)	30.35	114.17	27	21	(601007)	32.03	118.46	24
2	(000589)	26.35	106.42	123	22	(000550)	28.4	115.55	47
3	(000718)	43.54	125.19	32	23	(000799)	28.12	112.59	38
4	(002103)	30.16	120.1	211	24	(600502)	31.52	117.17	18
5	(600643)	31.14	121.29	25	25	(600690)	36.4	117	513
6	(600739)	41.48	123.25	46	26	(600806)	25.04	102.42	803
7	(600876)	34.46	113.4	324	27	(601628)	39.55	116.24	29
8	(000528)	22.48	108.19	541	28	(000625)	34.17	108.57	612
9	(000597)	41.48	123.25	680	29	(000813)	43.45	87.36	325
10	(000735)	20.02	110.2	32	30	(600004)	23.08	113.14	45
11	(002108)	38.02	114.3	34	31	(600609)	41.48	123.25	211
12	(600249)	22.48	108.19	321	32	(600859)	39.55	116.24	344
13	(600664)	45.44	126.36	409	33	(601919)	39.02	117.12	21
14	(600889)	32.03	118.46	12	34	(000573)	23.08	113.14	46
15	(000547)	26.05	119.18	23	35	(000651)	23.08	113.14	32
16	(000606)	36.38	101.48	411	36	(000886)	20.02	110.2	46
17	(000762)	29.39	91.08	781	37	(600622)	31.14	121.29	211
18	(002109)	34.17	108.57	316	38	(600738)	36.04	103.51	469
19	(600383)	30.4	104.04	412	39	(600863)	40.48	111.41	803
20	(600789)	36.4	117	36					

Table 4.2 correlative matrix for five stocks

k\j	1	2	3	4	5
1	1	0.48309	0.42045	0.4919	0.40621
2	0.48309	1	0.46836	0.53994	0.4141
3	0.42045	0.46836	1	0.46625	0.4444
4	0.4919	0.53994	0.46625	1	0.44414
5	0.40621	0.4141	0.4444	0.44414	1

Weis-Fogh wave power electricity generation device optimization parallel model

Quan Kuang
School of Energy & Power Engineering
Wuhan University of Technology
Wuhan, 430063, China
e-mail: 1053373184@qq.com

Xin Chen Shesheng Zhang
School of Logistics Engineering
Wuhan University of Technology
Wuhan, 430063, China
e-mail: 755973978@qq.com

***Abstract*—Based on the high lift generated by Weis-Fogh mechanism, two stage wave power electricity generation device is designed, a size objective optimization model of parallel computation is established, the formula of Green function is obtain, and hydrodynamics is calculated by using parallel computer.**

Keywords-wave power, electricity generation device; Weis-Fogh mechanism; optimization model; parallel calculation

I. INTRODUCTION

Wave energy is a clean and renewable energy. The use of wave power is one of the effective ways to solve the problem of energy shortage. From a practical point of view, the following problems of current wave power device is still not well solved : (a) a low energy conversion efficiency[1]; (b) low system reliability[2,3]. Therefore, further improve the conversion efficiency and stability of wave energy device is the key technology of the utilization of wave energy. Weis--Fogh device is a special mechanism to produce high lift.it produces lift rapidly[4]. In practical application, the main applications of Weis--Fogh effect are as follows: (a) apply to build marine propulsion[5,6];(b) apply to build the ship zero speed fin[7].The literature [8] showed a ideal of Weis-Fogh wave power device, and discussed its hydrodynamics. But the paper did not give a detailed mechanism scheme.

Aiming at the existing problems of power device and the current research situation of Weis-Fogh mechanism, we will discuss Weis-Fogh wave energy device.

II. DEVICE OVERVIEW

The Weis-Fogh wave energy device is shown in Fig.1. Its working process is as follows

（A）The lower box body is put in the water with a lifting wing. Acted by wave force, the wing is opened to certain angle, and then all wing moved upward to the up wall. During wing opening, the wing root is fixed on the low wall. During upward moving, the moving speed in any point on the wing is same.

（B）Wing drives the push rod up and down reciprocating movement ; the pusher also pushes the electric generator rotor up and down reciprocating movement that cut magnetic lines and eventually convert mechanical energy to electrical energy. The electrical energy is stored in a storage battery.

III. OPTIMIZATION MODEL

Consider regular waves, the wave horizontal speed is u=Acos(ωt) Suppose F is lifting force acted on wing, the m is mass of wing, we have wing moving equation in y direction:

$$m\ddot{y} = (1-b)F$$

Here b is control parameter, it controls wing moving direction and moving speed. In a wave period , the total work acted by force is:

$$W = \int_0^T (1-b)|F|dy = 2\int_0^{T/2} (1-b)F(t)dy \qquad (3\text{-}1)$$

Here, |F| is the absolute value of force, and is symmetric about t=T/2, We take F>0 if 0<t<T/2, take F>0, ifT/2<t<T, then F(t)=-F(T-t). By using difference method, above equation may be written as

$$W = 2\sum_{j=1}^{N} (1-b)F(t_j)\dot{y}(t_j)dt \qquad (3\text{-}2)$$

Here *dt* is time step length, N is points number in the domain[0,*T*/2]. Considering special case, when b=1, the work W is zero. when b=0, the work W is only done by lifting force. The lift force may be expressed by lifting coefficient C_L,

$$F = 0.5V_\infty^2 \rho c C_L ,$$

Here c wing chord length, ρ is water density,V_∞ is water coming velocity. Then the work is:

$$W = V_\infty^2 \rho c \sum_{j=1}^{N} (1-b)C_L(t_j)\dot{y}(t_j)dt \qquad (3\text{-}3)$$

Take $y=0.5\rho cz$ ， The wing moving equation may be written as:

$$\ddot{z} = (1-b)V_\infty^2 C_L$$

The two order difference formula is：

$$z_{j+1} = 2z_j - z_{j-1} + (1-b)V_\infty^2 \frac{C_L}{m}dt$$
$$z_0 = z_1 = 0 \qquad (3\text{-}4)$$

Form above formula, the work W is related with the control coefficient b of power mechanism. The wave force is calculated by below formula at the wing opening stage[6]:

$$i\frac{F}{\rho}e^{-\alpha\pi i} = \Omega^2 F_1 - u'F_3 + \Omega^2 F_2 - \Omega u F_4 + u^2 F_5 + i(\Omega^2 F_7 - \Omega u F_8 + u^2 F_9) \qquad (3\text{-}5)$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.9

here：

$$F_1 = -\mathrm{Re}\int_{-1}^{1} GKh'dt \qquad F_3 = -\mathrm{ReK}\int_{-1}^{1} \ln(t-b)Kh'dt$$

$$F_{22} = -0.5\int_{-1}^{1} G_\zeta^2 \,|\, z_\zeta' \,|^{-2}\, Kh'dt \qquad F_{42} = -K\int_{-1}^{1} \frac{G_\zeta}{t-b} \,|\, z_\zeta' \,|^{-2}\, Kh'dt$$

$$F_{52} = -0.5K\int_{-1}^{1} |\,(\mathrm{t}-\mathrm{b})\, z_\zeta' \,|^{-2}\, Kh'dt$$

$$F_2 = \mathrm{Re}\, F_{22} + \frac{F_{1\alpha}}{\pi} \qquad F_7 = \mathrm{Im}\, F_{22}$$

$$F_4 = \mathrm{Re}\, F_{42} + \frac{F_{3\alpha}}{\pi} \qquad F_8 = \mathrm{Im}\, F_{42}$$

$$F_5 = \mathrm{Re}\, F_{52} \qquad F_9 = \mathrm{Im}\, F_{52}$$

Here $(-F_7),(-F_8),(-F_9)$ are leading edge attractive force, its direction is from wing tip to the wing root., see Fig.2.

After end opening, wing moves upward, the water force is calculated by the formula below:

$$F = \int_c |\int_c \sigma(Q)\frac{dG}{dn}dl\,|^2 \vec{n}ds \tag{3-6}$$

Her $\vec{n}$ is wing normal vector, G is Green function determined by below boundary problem:

$$\frac{\partial^2 \mathrm{G}}{\partial x^2} + \frac{\partial^2 \mathrm{G}}{\partial y^2} = \delta(x-\xi)\delta(y-\eta) \tag{3-7}$$

$$\frac{\partial \mathrm{G}}{\partial y} + \mathrm{kG} = 0,\ y = 0$$

$$|\,\mathrm{G}| < \mathrm{C} \qquad |\frac{\partial \mathrm{G}}{\partial x} \pm j\mathrm{kG}| < \frac{\mathrm{C}}{|\mathrm{x}|} \qquad x \to \mp\infty$$

Here point source strength $\sigma(\mathrm{P})$ calculated by below integral equation：

$$n_y = \frac{\sigma(P)}{2}\mathrm{n} + \frac{1}{2\pi}\int_S \sigma(Q)\frac{\partial}{\partial n}G(P,Q)ds \tag{3-8}$$

Where P is source point, Q is a point on the wetted wing surface of x section, and G(P,Q) is Green function given in above section. σ is the source strength on the boundary S of floating body wetted surface. n_y is normal vector n projecting on y direction.

According to the above discuss, the absorb wave energy optimization problems of Weis-Fogh wave power generation device is

$$\max W(c,\alpha) = \sum_{j=1}^{N} (1-b)F(t_j)\dot{y}(t_j)dt \tag{3-9}$$

here y is determined by below problem:

$$m\ddot{y} = (1-b)F$$

$$\mathrm{y}(0) = 0$$

$$\dot{\mathrm{y}}(0) = 0$$

Where lifting force F is related wing chord length and maximum opening angle alpha. So that above calculating work W is nonlinear problem. It is difficult to find the analytical solution, so numerical algorithm is used to compute the optimal solution

IV. NUMERICAL CALCULATE

Above discuss tall us that the Green function is calculated by boundary problem, and the wing lifting height is the solution of tow order nonlinear different problem. The Green function may be calculated by below formula:

$$\mathrm{G} = \frac{1}{2\pi}\mathrm{Re}\{\ln\frac{Y-2k\eta}{Y}\} - j\,\mathrm{Re}\{e^{-Y}\) - \frac{1}{\pi}\mathrm{Re}[H(Y)] \tag{4-1}$$

Here Y=k[y+η-j(x-ξ)], and function H(Y) is defined as

$$H(\mathrm{Y}) = -\exp(-Y)[C_E + \ln(Y) + \sum_{n=1}^{\infty}\frac{Y^n}{n!n}] \tag{4-2}$$

C_E=0.577.., is Euler constant

σ is the source strength on the boundary c of hydrofoil wetted surface, and solved from below equation:

$$n_y = \frac{\sigma(P)}{2}\mathrm{n} + \frac{1}{2\pi}\int_c \sigma(Q)\frac{\partial}{\partial n}G(P,Q)ds \tag{4-3}$$

From above formula, the optimization computation need long CPU time. Therefore, this paper takes parallel computer to shorted wall time. The parallel calculation steps are:

(1) choose P processor, given airfoil section chord length and opening angle alpha, and divided opening angle domain into P member area;

(2) In each sub region arranged a processor computes values of Green function;

(3) data is transmitted to the master processor, calculate source strength;

(4) the source strength is transmitted to other slave processors, calculate the force F and W;

(5) the W is transmitted to the master processor, seeking out the optimal chord length and opening angle alpha.

Consider wing is a flat plate with unit length, the depth in the water is 0.5, choose P=20. The H function of Green function is calculated and shown in Fig.3. In the figure, the horizontal axis is real part of H function, the vertical axis is imaginary part of H function. The parallel computation efficiency up to 0.84, and speedup is 16.8. The work efficiency is over 60%.

V. CONCLUSION

Wave energy conversion efficiency is the main factor of wave power generation device. The Weis-Fogh wave power generation device may improve such conversion efficiency. The hydrodynamics can be calculated by using Weis-Fogh theory

and Green function theory. The parallel computer is chosen to shorted wall time.

ACKNOWLEDGMENT

The paper is financially supported by Students innovation a nd entrepreneurship training program,Wuhan university of techn ology, China.(No.2014104970500）.

REFERENCES

[1] Masaaki Imari. Yukihisa Washio, Hirotaka Osawa. Development of an offshore floating type wave power energy converter system " mighty whale" [J].Science &Technology in Japan. 1997,60(15): 81-84.

[2] Xin Chen, Shengping Jin, Shesheng Zhang, Dan Li. A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil[J], Journal of Algorithms & Computational Technology,2015,Vol. 9 No. 2, 154-161.

[3] Ying Z, Zaili D, Chen dong W. Modeling of wave energy absorption based on BP Neural network for underwater vehicles: Intelligent Control and Automation, 2008.WCICA 2008. 7th World Congresson, 2008[C].25-27 June 2008.

[4] Qiu Zhizhen, Xie Nenggang. The reversed thinking and outlet ofanimal-motion bionics ［C］ //Proceedings of the 6th international conference on frontiers of design and manufacturing, Xi'an, China, 2004： 477-478

[5] Tsutahara M,Kimura T. An Application of the Weis-Fogh Mchaninsm to Ship Propulsion, Journal of Fluid Engineering（ASME) , 1987, 109(2), 107-113

[6] Zhang Shesheng, Wu Xiuheng, Wang Xianfu. Cascade Weis-Fogh hydrofoil propulsion mechanism study on hydrodynamic performance of [J]. shipbuilding of China, 1998, S1:42-54.

[7] Jin Hongzhang, Qi Zhigang, Luo Yanming, Gong Jin. Study on [J]. Journal of system simulation of zero speed fin stabilizer Weis-Fogh mechanism model based on 2007,17:4079-4081.

[8] Xin Chen, Quan Kuang, Yang Li, Songbo Wang, Yunling Ye, Shesheng Zhang. Weis-Fogh Mechanism Mathematic Model of Wave Power Generation Device[J]. Open Journal of Fluid Dynamics, 2014, 4:373-378

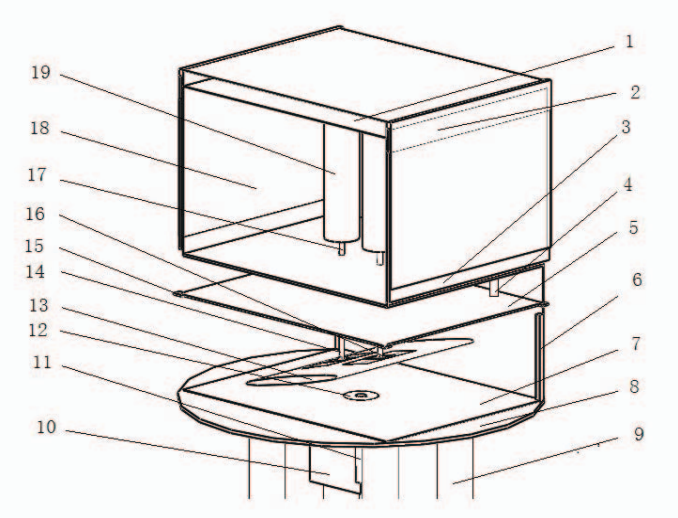


Figure 1. Device for local section view

1, the upper box body; 2, the limiting case; 3, the baffle plate; 4, the collision switch; 5, the sliding plate; 6, limiting column; 7, the lower box body; 8, a large disk; 9, jacket; 10, the rudder plate; the 11, connecting shaft; 12, the wing plate; 13, a front push rod; 14, after the push rod; 15, sleeve, 16, middle layer box body; 17, linear generator; 18, the self-locking claw, 19, small disc.

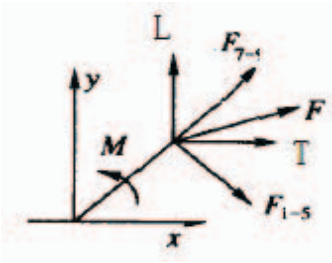


Figure 2. Force and direction

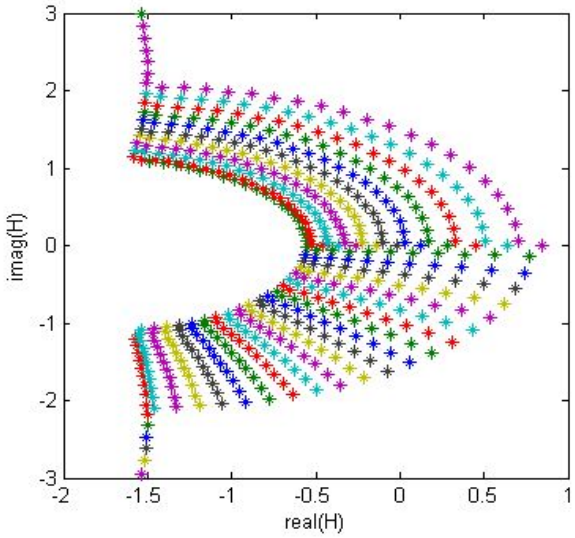


Figure 3. The real art of H function varied with imaginary parts

The lager ship fluid-solid coupling parallel algorithm based on VOSS mapping theory

Yuguang Li
Department of Statistics, Wuhan University of Technology , Wuhan 430063, China
LiYuguang@qq.com

Xin Chen Shesheng Zhang
Wuhan University of Technology, , Wuhan 430063, China
sheshengz@qq.com

***Abstract*—The fluid-solid coupling hydrodynamics parallel model of large ship is built based on VOSS mapping theory. The VOSS algorithm of large ship's wave is given, and the control equation and initial boundary conditions are derived from fluid hydrodynamic theory and solid elastic theory. The numerical results show the VOSS transfer of ship fluid-solid coupling wave is not same as coming see wave.**

Keywords-component; parallel computer, fluid solid coupling, VOSS theory

I. INTRODUCTION

The ship will be swayed on the wave water, and reduce efficiency of propeller. When the wave height reaches a certain value, the wave energy will make ship hull deformation, or even crash. Therefore, understanding the ship wave characteristics can guarantee the quality of ship navigation.

We can built wave hydrodynamics model based on wave data measured from see wave, or wave spectrum[1]. By using Green function, linear wave characteristic model is often used to show ship hull deformation on the wave water.[2]. According to the water viscosity and wind speed, the wave energy will be varied with time, the wave length model is built to analysis wave energy calculated from wave length [3]. By using VOSS theory[4], the VOSS mapping model is built to analysis wave energy[5], and paper[5] showed VOSS mapping wave power spectrum for different wind speed.

This paper will consider the fluid-solid coupling problem by using VOSS theory. Section II will introduce basic VOSS theory; section III will show ship fluid-solid coupling initial boundary problem; section IV will analysis ship fluid-solid coupling wave by using VOSS parallel calculation method, section V will discuss numerical results..

II. VOSS THEORY OF WAVE

Suppose set S={S(n): S(n)∈I,n=0,1,2,⋯,N-1} is a sequence of wave height measured from wave water, In the set S, A1 is ,minimum value, AL is maximum value. Let set I={A1，A2,⋯,Ak} is all value of set S. It is find that the number of element in the set I is K . For any value b∈I,, let

$$u(b,n)=\begin{cases}1 & S(n)=b\\ 0 & S(n)\neq b\end{cases} \qquad (2,1)$$

We find u(b,n) is also a sequence produced from set S. Take Fourier transfer:

$$U(b,k)=\sum_{n=0}^{N-1}u(b,n)\exp(-j\frac{2nk\pi}{N}) \qquad (2,2)$$
$$k=0,1,2,..,N-1,b\in I$$

We know b∈I，and there is K element in the set I. So that there is K series Fourier transfer. When N=2^m, the series can be mapped by using fast Fourier transfer(FFT). The amount of calculation is K*N*log(2N). On other case, FFT can't be used to transfer u(b,n), and they must be calculated one by one, The amount of calculation is K*N*N. We know u(b,n)=0 or 1, so that U(b,k) may be rewritten as:

$$U(b,k)=\sum_{u(b,n)\neq 0}^{1}\exp(-j\frac{2nk\pi}{N}) \qquad (2,3)$$
$$k=0,1,2,..,N-1,b\in I$$

When k=0, we have

$$U(b,0)=\sum_{u(b,n)\neq 0}^{1}1=N(u(b,n)\neq 0) \qquad (2,4)$$

When k=N/2, we have

$$U(b,0)=\sum_{u(b,n)\neq 0}^{1}\exp(-jn\pi)=\sum_{u(b,n)\neq 0}^{1}(-1)^n\neq 0)$$

Power spectral sequence {P(k)} is define as：

$$P(k)=\sum_{b\in 1}|U(b,k)|^2$$
$$k=0,1,...,N-1$$

And total power is

$$E=\frac{1}{N}\sum_{k=0}^{N-1}P(k) \qquad (2.5)$$

Consider the wave depression as:

$$f(x)=\begin{cases}\frac{2x}{a} & 0\le x<\frac{a}{2}\\ 2-\frac{2x}{a} & \frac{a}{2}\le x<a\\ 0 & a\le x<a+b\end{cases}$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.10

And $f(x)=f(x+T) \quad T=a+b$

The value of function f(x) are:

R1=[0, 1/k)， R2=[1/k, 2/k)，…，Rl=[(l-1)/k, l/k)，…，Rk=[(k-1)/k,1)

The boundary of sub-domain are:

$$c_1,\ c_1+\frac{a}{2k},\ ...,\ \ c_1+\frac{(l-1)a}{2k},\ c_1+\frac{la}{2k},$$
$$...,\ c_1+\frac{(k-1)a}{2k},\ \ c_1+\frac{a}{2},\ \ c_1+\frac{a}{2}+\frac{a}{2k},$$
$$...,\ \ c_1+\frac{a}{2}+\frac{(l-1)a}{2k},\ \ c_1+\frac{a}{2}+\frac{la}{2k},$$
$$...,\ \ c_1+\frac{a}{2}+\frac{(k-1)a}{2k},\ \ c_1+a,\ \ c_1+a+b,$$
$$...,\ \ s,s+\frac{a}{2k},\ \ ...,s+\frac{(l-1)a}{2k},\ \ s+\frac{la}{2k},$$
$$...,\ \ s+\frac{(k-1)a}{2k},\ \ s+\frac{a}{2},\ \ s+\frac{a}{2}+\frac{a}{2k},$$
$$...,\ \ s+\frac{a}{2}+\frac{(l-1)a}{2k},\ \ s+\frac{a}{2}+\frac{la}{2k},$$
$$...,\ \ \ s+\frac{a}{2}+\frac{(k-1)a}{2k},\ \ s+a,s+a+b,,...d_1,$$
$$d_1+\frac{a}{2k},\ \ ...,d_1+\frac{(l-1)a}{2k},\ \ d_1+\frac{la}{2k},$$
$$...,\ \ \ d_1+\frac{(k-1)a}{2k},\ \ d_1+\frac{a}{2},d_1+\frac{a}{2}+\frac{a}{2k},$$
$$...,\ \ d_1+\frac{a}{2}+\frac{(l-1)a}{2k},\ \ d_1+\frac{a}{2}+\frac{la}{2k},$$
$$...,\ \ d_1+\frac{a}{2}+\frac{(k-1)a}{2k},\ \ d_1+a,d_1+a+b.$$

here

$$2\le l\le k, c1\le s\le d1, c_1=\left[\frac{c}{T}\right],$$
$$d_1=\left[\frac{d}{T}\right], T=a+b$$

On the domain Rl=[(l-1)/k, l/k)， we have:value of xi as below:

$$h_1(s,l)+1,\ \ h_1(s,l)+2,\ \ ...,\ \ h_2(s,l),$$
$$j_1(s,l)+1,\ \ j_1(s,l)+2,\ \ ...,\ \ j_2(s,l)$$

Here:

$$h_1(s,l)=less(\frac{\left(s+\frac{(l-1)a}{2k}-c\right)N}{d-c}),$$
$$h_2(s,l)=less(\frac{(s+\frac{la}{2k}-c)N}{d-c})$$
$$h_1(s,l)=less(\frac{\left(s+a-\frac{la}{2k}-c\right)N}{d-c}),$$
$$j_2(s,l)=less(\frac{(s+a-\frac{(l-1)a}{2k}-c)N}{d-c})$$
$$2\le l\le k,\ \ c_1\le s\le d_1,\ \ c_1=\left[\frac{c}{T}\right],\ \ d_1=\left[\frac{d}{T}\right],$$
$$T=a+b$$

We get the U value:

$$U_l(p)=\sum_{s=c_1}^{d_1}(X_p^{h_1(s,l)+1}\times\frac{1-X_p^{h_2(s,l)-h_1(s,l)}}{1-X_p}$$
$$+X_p^{j_1(s)+1}\times\frac{1-X_p^{j_2(s,l)-j_1(s,l)}}{1-X_p})$$

Here

$$h_1(s,l)=less(\frac{\left(s+\frac{(l-1)a}{2k}-c\right)N}{d-c}),$$
$$h_2(s,l)=less(\frac{(s+\frac{la}{2k}-c)N}{d-c})$$
$$h_1(s,l)=less(\frac{\left(s+a-\frac{la}{2k}-c\right)N}{d-c}),$$
$$j_2(s,l)=less(\frac{(s+a-\frac{(l-1)a}{2k}-c)N}{d-c})$$
$$2\le l\le k,\ \ c_1\le s\le d_1,\ \ c_1=\left[\frac{c}{T}\right],\ \ d_1=\left[\frac{d}{T}\right],$$
$$T=a+b$$

III. SHIP FLUID-SOLID COUPLING WAVE

From paper[6], the control equation of ship fluid-solid coupling wave is:

$$(\rho + \frac{\rho_m}{S_x})\frac{\partial^2 w}{\partial^2 t} = G\frac{\partial^2 w}{\partial^2 x} - \rho g + \rho_w g\frac{s_0}{S_x} + \rho_w g\frac{s(z-v}{S_x} \quad (3.1)$$

Here w is the average value of displacement on the x=x section (or x section), ρ m is the added mass of the x section, ρ w is the water density, Sx is the solid body area of x section, s0 is the area below z0 , z0 is the coordinate of water surface when water is in the static state., s(z)+ s0 is the area below z, and s(z0)=0, s(z)>0 when z > z0, s(z)<0 when z<z0, g =9.81m/s2 is the acceleration of gravity. ρ m wtt is added mass force, is the water force acted on ship hull buoyancy, z=z-0 and (z-w) are the height of water surface when hull displacement is zero and w, respectively. If displacement w<0, (z-w)>z. The initial conditions is:

$$w(t=0) = w_0 \qquad w'(t=0) = w'_0 \quad (3.2)$$

The boundary conditions are:

$$w(x=0) = w_a \qquad w(x=L) = w_L \quad (3.3)$$

Where L is the length of ship. In the above motion equation, the added mass is calculated using the two dimension fluid dynamics theory.

If added mass and the section area are constants, and area s(z-w)=d0(z-w), d0 is constant, K is wave number, ω is wave cycle frequency, and， let

$$a^2 = G/(\rho + \frac{\rho_m}{S_x}) \qquad \rho_1 = [-\rho g + \rho_w g\frac{s_0}{S_x}]$$

$$b = \rho_w g\frac{d_0}{S_x} \qquad \rho_3 = -\rho_w g\frac{1}{S_x} \qquad H_1 = bH \quad (3.4)$$

The ship control equation becomes:

$$\frac{\partial^2 w}{\partial^2 t} = a^2\frac{\partial^2 w}{\partial^2 x} + \rho_1 + H_1\sin(Kx - tL_1\omega) - bw \quad (3.5)$$

Below section will discuss calculation of above equation.

IV. VOSS PARALLEL CALCULATION

Let h is space step length, τ is time step length, w(k)j=w(xj,tk) is deformation function w at point (xj,tk), the ship control equation may be written as:

$$w_i^{(k+1)} = 2w_i^{(k)} - w_i^{(k-1)} + \frac{a\tau^2}{h^2}(w_{i+1}^{(k)} - 2w_i^{(k)} + w_{i-1}^{(k)})$$
$$+\tau^2 f_i^{*(k)} - b\tau^2 w_i^{(k)}$$
$$w_0^{(k)} = \frac{1}{3}(4w_1^{(k)} - w_2^{(k)})$$
$$w_N^{(k)} = \frac{1}{3}(4w_{N-1}^{(k)} - w_{N-2}^{(k)})$$

$$w_i^{(0)} = \varphi(x_i) \qquad w_i^{(1)} = w_i^{(0)} + \tau\psi(x_i) \quad (4.1)$$

Here

$$f^*(x,t) = \rho_1 + H_1\sin(Kx - \omega t)$$

Take k=1, we have,

$$w_i^{(2)} = 2w_i^{(1)} - w_i^{(0)} + \frac{a\tau^2}{h^2}(w_{i+1}^{(1)} - 2w_i^{(1)} + w_{i-1}^{(1)})$$
$$+\tau^2 f_i^{*(1)} - b\tau^2 w_i^{(1)}$$
$$w_0^{(1)} = \frac{1}{3}(4w_1^{(1)} - w_2^{(1)})$$
$$w_N^{(1)} = \frac{1}{3}(4w_{N-1}^{(1)} - w_{N-2}^{(1)})$$

Take k=2, we have,

$$w_i^{(3)} = 2w_i^{(2)} - w_i^{(1)} + \frac{a\tau^2}{h^2}(w_{i+1}^{(2)} - 2w_i^{(2)} + w_{i-1}^{(2)})$$
$$+\tau^2 f_i^{*(2)} - b\tau^2 w_i^{(2)}$$
$$w_0^{(2)} = \frac{1}{3}(4w_1^{(2)} - w_2^{(2)})$$
$$w_N^{(2)} = \frac{1}{3}(4w_{N-1}^{(2)} - w_{N-2}^{(2)})$$

The VOSS parallel calculation steps of ship fluid-solid coupling are:

1) give h, τ and all parameters in the control equation;
2) choose P processors, and divide numerical domain into P sub-domain;
3) calculate ship deformation w by using parallel algorithm;
4) calculate 0-1 sequence u(b,n);
5) calculate mapping sequence U(b,k)
6) calculate power spectral sequence {P(k)}
7) calculate total power

The power spectral sequence is shown in the Fig.1. In the figure, the parameters are, K=1000, N=10000, L=100m, wind speed V=1m/s, P=32. There are about 20 peeks in the

figure, the interval about 50 for k. The peek is lowest in the middle of figure, and highest in two sides. The parallel speed up is 25.32

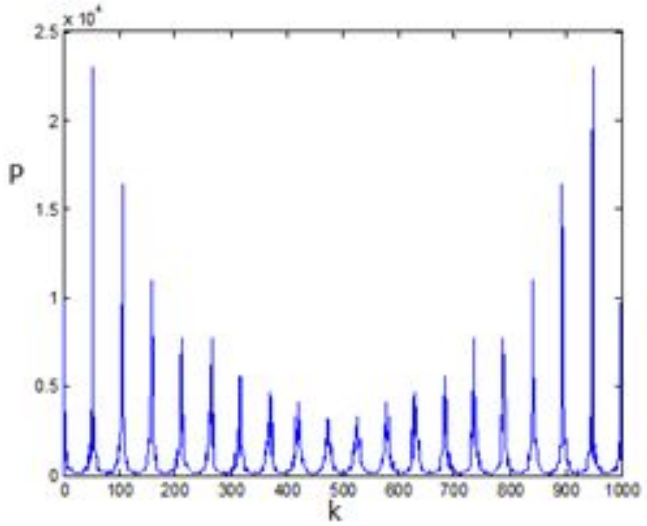


Fig.1 power spectral sequence for wind speed V=1 m/s

V. CONCLUSIONS

The big ship deformation wave is not same as water wave. The first one is the results calculated from fluid-solid coupling initial boundary problem. The parallel VOSS mapping is a fast analysis algorithm to study ship deformation in the see water.

ACKNOWLEDGMENT

The paper is financially supported by China national natural science foundation (No.51139005),

REFERENCES

[1] Xin Chen, Mengyu Li, and Shesheng Zhang，A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil［J］, Journal of Algorithms & Computational Technology,2014,Vol. 8 No. 3，249-266.

[2] [2] Ma Jie, Peng Tian, demobilization, numerical simulation of ocean wave and the simulation of [J], Huazhong University of science and technology, 4 (2004), 63-65.

[3] [3] Li Mengguo, Wang Zhenglin, Jiang Decai, research and development of mathematical models of wave transformation [J] offshore, ocean engineering, 2002 (4), 43-57

[4] [4] Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78‑94.

[5] [5] Li Mengyu Zhang Shesheng, Fast calculation of ship encountering wave VOSS mapping[J], Journal of ChangJiang University, 2013（4），1-3.

[6] [6] Xin Chen, XinCong Zhou, Shesheng Zhang，Dan Li，A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil［J］, Journal of Algorithms & Computational Technology,2015,Vol. 9 No. 2，163-175. ..

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

ZDLC: Towards Automated Software Construction and Migration

S. Khaddaj, Y.K. Makoondlall
School of Computing and Information Systems, Kingston University, Kingston upon Thames, UK
e-mail: s.khaddaj@kingston.ac.uk

B. Makoond, S. Chivukula, K Kethan
Cognizant Technology Solutions
London SW1Y 4SP, UK
e-mail: B.Makoond@cognizant.com

***Abstract*— The construction of modern software projects with a multitude of quality requirements is a challenging and complex task particularly for large scale systems. Underestimating this complexity may result in project failures at worst or the delivery of poor quality products that are plagued with software defects at best. However, introducing scientific rigor at every step of the software development life cycle (SDLC) can reduce defect leakage through the different stages. While this can improve quality considerably it might affect upfront cost and speed of delivery. Thus, there is a clear need to introduce automation in software production and defect tracking. The Zero Deviation Life Cycle (ZDLC) framework addresses this conflict i.e. quality vs. cost, by proposing a methodology that can be applied not only to new projects but also to legacy applications.**

Keywords- Zero Deviation Life Cycle (ZDLC), software quality, automation of software construction, legacy systems .

I. INTRODUCTION

Large-scale software projects failure rates are considerably higher than many other engineering disciplines mainly due to their complexity and diversity. There are many examples of major projects that have been plagued with delays and high running cost as a result of poor requirements capture and/or inadequate construction and testing [1]. On the other hand market pressures have resulted in the premature delivery of the many software products which later found to be plagued with software defects [2]. Thus, there is an urgent need to introduce scientific rigor at every step of the software development life cycle (SDLC) in order to detect, track and stop defects leakage. However, introducing verification and validation processes at every stage of the SDLC is very time consuming, therefore automation of the processes is becoming essential. The approach adopted in this paper is based around the ZDLC (Zero Deviation Life Cycle) philosophy [3], which aims to tackle many of the complex problems faced in today's software industries.

Thus, to address these problems we consider automation techniques and tools necessary not only for the development of the different SDLC stages but also for minimizing defects leakage at every stage, and guaranteeing quality through continuous verification and validation. This should start with business processes and requirement engineering, through design and architecture, and goes on to implementation and testing.

In reality Software Quality Assurance (QA) should be at the core of any system development [4], [5], and one of the novel aspects of ZDLC is that QA is considered as early as possible in the SDLC in order to reduce software defects leakage through the different SDLC stages, for example, requirement engineering covers not only requirements capture but also validating the correctness of the process. However, apart from software quality it is important to deliver the projects on schedule and on budget. Some of the cancelled projects were simply called off as they exceeded their initial budget and were behind schedule. Thus, automation of many of the software engineering processes, particularly quality assurance is very important.

The same problems and issues arise when dealing with legacy applications. Legacy systems are referred to systems that are increasingly difficult to modify and evolve to meet new and constantly changing business requirements, and make use of the latest technological advances. In fact there is a constant need to find solutions to problems of legacy software. Businesses need to keep their information and communication technology up to date and in line with current business requirements, security threats and technological advances. This requires migration of legacy system to modern technology that might include changing the software infrastructure. The changes in architecture mean that the software application needs to be reprogrammed or ported in different programming languages and paradigms to fully utilize the capability of modern technology.

Legacy software migration is an immensely large and complex task. The task of overseeing an IT migration project or software projects to completion is a complicated one, which is full of pitfalls. It requires plenty of tools to re-design and reverse engineer existing systems. However, due to the huge risk and cost associated with software migration, there is a massive number of applications that still rely on technologies and programming languages that were developed as early as the 1960s. In fact a large number of financials applications and systems are still using COBOL [6]. Thus, a step by step strategy based on continuous validation and verification, as well as automation will also be of great benefits to software migration. Whilst it is true that software failures or rather software malfunction is likely to be around for as long software will be, all the key stakeholders involved in a project need to ensure that the project is completed with as little defects as possible.

This paper starts with a brief discussion of the background of the Zero Deviation Life Cycle (ZDLC)

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.11

approach. Then, automation techniques based on natural language processing are considered. This is followed by a presentation of the main architecture of the proposed framework. Finally, a summary of the finding is presented.

II. Zero Deviation Life Cycle (ZDLC)

The Zero Deviation Life Cycle (ZDLC) is an approach to transform any Software Development Life Cycle (SDLC) or software legacy migration into a Software Value Chain, which means that at any point in time ZDLC enforces value creation for the industries [7]. ZDLC has been devised specifically to ensure minimum deviation from the requirements and to reduce defect leakages between the different phases of the SDLC. ZDLC sustains Value Creation by embedding capabilities such as Automation, Formal Methods, Advanced Simulation and Artificial Intelligence all packaged into an intuitive and simple tool set that employs the concept of gamification and advanced usability techniques to increase ease of use and adoption [8]. ZDLC is supported by a number of tools and software products, designed to practice the core principles of Quality Engineering throughout the Software Life Cycle.

Time, accuracy and quality within the SDLC can be improved, but we need to rethink the use and the dynamics of classical Software Engineering tools. New techniques are required and ZDLC is a framework that provides an approach to merge classical tools with scientific techniques so as to augment accuracy, productivity and quality of delivery whilst reducing effort. This led to the evolution of Software Development Life Cycle to the Software Value Chain [8].

ZDLC tools can help to reduce defect leakages between the different phases of the SDLC and also minimize deviation from the requirements. However there are no tools which automate the process of translating the requirements into software models. The aim of this paper is to propose a framework which automates the translation of natural language requirements, and legacy applications, into design artefacts. Thus, it is mainly concerned with expanding two ZDLC components namely the Requirements Modeling System (RMS) and the Systemic Defect Profiler (SDP). The next section exposes the textual analysis for the automation of requirements capture, for both new projects and legacy applications into software artefacts.

III. Textual Analysis Automation Techniques

There are a number of management techniques which textually analyses the requirements with the aim to assist in the design process [9], [10]. They employ Natural Language Processing (NLP) techniques to parse the requirements and to attempt to automatically generate models, which can be used in the design process. These models are typically UML artefacts, which can be used as the blue prints for software construction [9]. These include AbstFinder [12], which is a prototype tool that can be used to find abstractions in natural language text.

Amongst all the published work, the one accomplished by Kof [12] had demonstrated that NLP was mature enough to be applied for Requirement Engineering as the aim of NLP is not to understand the requirements, but to extract concepts from the requirements. In the approach proposed by Kof, the system engineer is solicited to validate the parses at various stages and can thus ensure that the requirement documents, containing all the key requirements vital for the software, are written down.

This research aims to produce a framework which will model all the three categories of requirements, coupled with a learning system which will allow the parses to grow more accurate over time. In so doing, the approach will provide a comprehensive tool to model the natural language requirements, better adapted for modelling distributed systems. The outputs from the tool will be three sets of design artefacts, drawn from one set of requirements which will then be used to elaborate the architecture of the software. The resulting enterprise architecture and the technical architecture are more likely going to be aligned with the requirements, thus reducing requirement defects.

IV. Proposed Framework

A. Development approaches

Traditionally, the main software development phases include the requirements phase, the design phase, the implementation phase and the testing phase. It may seem mostly describing the waterfall methodology but even for other traditional approaches or more modern agile approaches, the requirements will have to be ready before the Architectural Design can be started. Figure 2 illustrates the key stages in this approach which covers traditional SDLC stages.

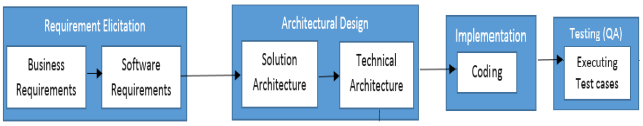


Figure 2: Traditional SDLC stages

While the migration of legacy applications follows a similar approach, the main difference is the early stages which require thorough assessment of existing code, documentation, parsing of code, etc…, in order to clearly define the requirements before proceeding to the reverse engineering phase (re-design) which is then followed by implementation and testing (Figure 3). We refer to this approach as Software Migration Life Cycle (SMLC).

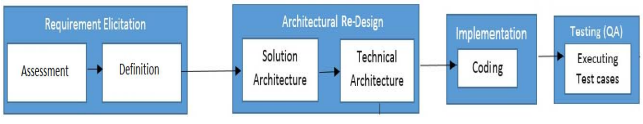


Figure 3: SMLC for legacy software migration

Overall, during the requirements phase, the client expresses the business needs which are collated as the

business requirements. From these business requirements the software requirements are derived, describing what the software needs to accomplish in order to deliver on the business requirements. For legacy systems this is defined by assessing existing code. Afterwards the architectural design is elaborated. The first step is designing the solutions architecture, which can be considered as the architecture of the solution as a whole. The technical architecture defines the low-level architecture of the program. For example the technical architect may opt for the MVC (Model View Controller) architecture at an application level. Usually the programming language is selected at this stage. During the implementation phase, the solution is coded and then sent to the Quality Assurance (QA) team to be tested.

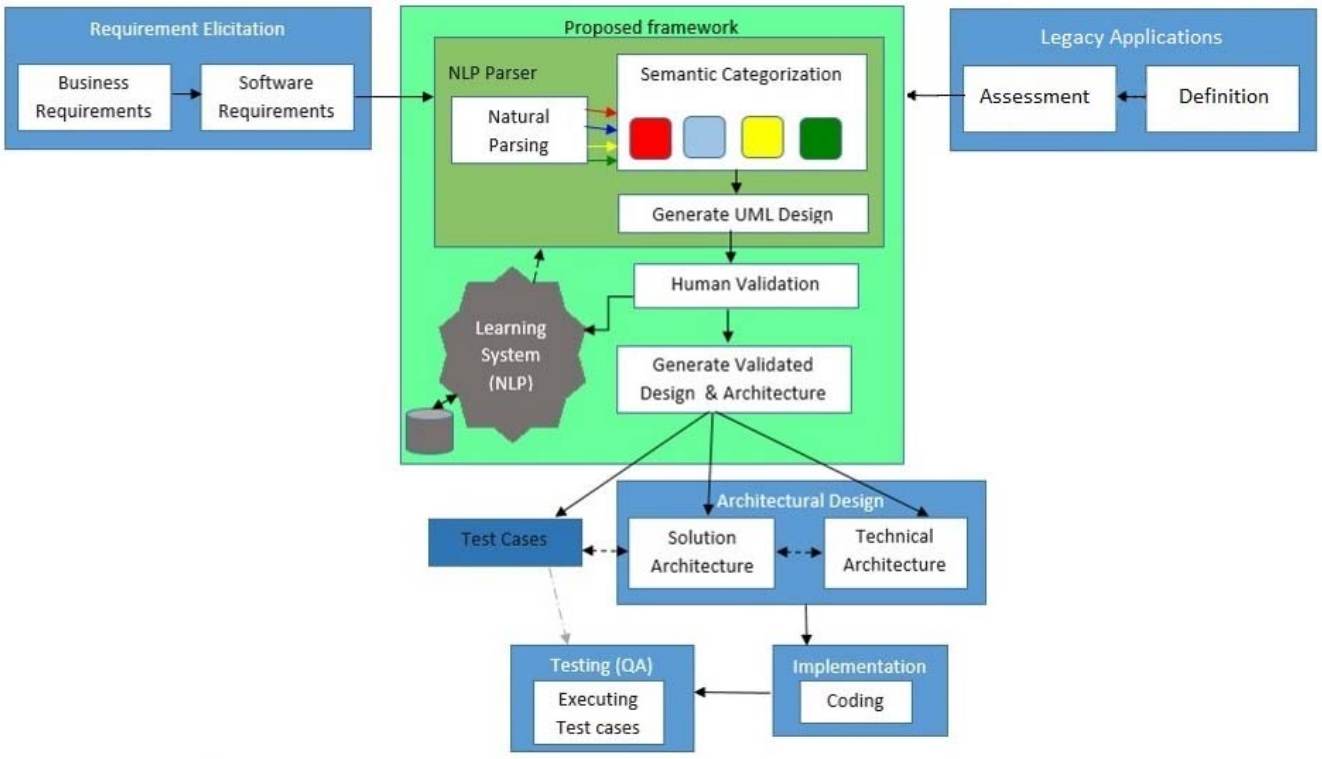


Figure 4 The proposed Approach

B. Proposed framework

The aim of this research is to propose a framework which helps the interpretation of the requirements and automation of the design process so as to reduce the number of defects transpiring from the requirement phase into the later stages of the SDLC. Moreover, the framework applies to both new and legacy applications. Figure 4 illustrates how the proposed framework alters the SDLC and there are five additional stages involved. In addition one stage is used solely for Software Migration Life Cycle (SMLC). Firstly, the Natural Language parser extracts key concepts from the requirements and secondly these concepts are categorized into four main categories. Thirdly, these categorized requirements are used to generate UML models which are presented to the user through the user interface. Fourthly, the user is given the possibility to modify the UML models and these modifications are captured by the learning system. Finally, the user validates the UML design and the Framework and then creates the finalized models. In addition the framework includes stages related to legacy applications, assessment, definition, and parsing before the categorization of business requirements of the legacy application migration. For legacy systems the parser will be trained using existing code but overall still following the same steps specified in the framework. The non-functional requirements will not be directly processed by the framework, but may be used at a later stage to extract the test cases.

The applicability to new projects:

The software requirements are fed to the NLP (Natural Language Processing) parser of the framework. From the requirements written in a Natural Language, the parser will identify sentences which pertain to four categories of requirements namely:

- Requirements pertaining to flow of data. (Red)
- Requirements pertaining to a process or process flow. (Blue)

- Requirements pertaining to communication sequences. (Yellow)
- Non-functional requirements. (Green)

The next step is the semantic categorization of the requirements. The requirements pertaining to the flow of data or data attributes are usually modelled with an Entity Relationship Diagram (ERD) or a class diagram. The requirements describing how a process should work, are usually modelled with a Data Flow Diagram or illustrated with a flowchart. The communication sequences within a system are usually modelled using a Sequence diagram. The non-functional requirements are not modelled, but can contain information describing certain quality attributes of the software.

The framework will group all the requirements belonging to the same category together (data, process, communication, or non-functional requirements). Those classified requirements are used to generate the models. The framework will analyse the categories individually. An algorithm will loop through all the requirements pertaining to data and then identify the relevant entities or classes. From these classes, the framework will identify the attributes of a class. All this information will then be used to generate the Entity Relationship diagram or the class diagram to describe the flow of data for the system.

The applicability to legacy applications:

After assessment and definition of the legacy migration and its business domain, the framework will parse the legacy code to extract the process flow, including the business logic and produce a Data Flow diagram or a flow chart. The framework will also extract the key variables and statements pertaining to communication and produce a use case and sequence diagram. Overall, similar steps, with some variations, to the new projects will be followed for legacy systems migrations.

All the artefacts produced by the framework are then presented to the user through a User Interface where the models can be viewed in turn. The user will also have the possibility to edit, change and modify the generated models. All the changes made by the user will be captured by the framework's learning system. The learning system is connected to a database, which will be used to store the changes made by the user and other parameters. The NLP parser will then use the data stored by the learning system so that future parses are more accurate.

When the user is satisfied with the models, the framework will then generate the models to be used for the design of the solution. The solutions architect will still have to validate the design artefacts outputted by the framework. The process has been partially automated and the technical architecture and the solution architecture are more likely going to be aligned, as they are modelled from the same requirement set. The models produced by the framework and the non-functional requirements can also be used to derive the test cases for the system.

V. CONCLUSION AND FUTURE WORK

In this work a framework aims to automate the translation of Natural Language requirements, and legacy applications into design artefacts. The framework includes a learning system and reinforcement model, in order to ensure that the parsing capacity of the tool evolves over time. It provides a comprehensive tool to model data requirements, process requirements and communication requirements. The automation processes ensure that design architecture and the code do not diverge from the requirements. The proposed framework can be used in conjunction with the toolset available for ZDLC, after the development and implementation of the different algorithms to parse the Natural Language requirements and the legacy code. The learning system will also be coded, tested and evaluated. Once the main components are developed a number of case studies will be evaluated.

REFERENCES

[1] Charette, Robert N. "Why software fails." IEEE spectrum 42, no. 9 2005.

[2] Zhivich, Michael, and Robert K. Cunningham. "The Real Cost of Software Errors." IEEE Security & Privacy Magazine 7.2 (2009): pp 87–90, 2012.

[3] Makoond, B. Elias, A., Talbot, S.R., Khaddaj, S., Franczuk, S., "ZDLC-Based Modelling and Simulation of Enterprise Systems", pp. 237-243, High Performance Computing and Communications, IEEE publications, 2014.

[4] Khaddaj, Souheil, and Gerard Horgan. "The evaluation of software quality factors in very large information systems." Electronic Journal of Information Systems Evaluation 7.1, pp 43-48, 2004.

[5] G. Horgan, S. A. Khaddaj, "Use of an adaptable quality model approach in a production support environment" in 'Journal of Systems and Software', 82(4), Elsevier, April, pp. 730-738, 2009.

[6] Valerio Cosentino, JordiCabot, Patrick Albert, Philippe Bauquel, Jacques Perronnet. Extracting Business Rules from COBOL: A Model-Based Framework. Working Conference on Reverse Engineering, Koblenz, Germany, 2013.

[7] http://ovum.com/2013/02/04/zero-deviation-lifecycle-gives-requirements-engineering-and-software-modeling-a-refresh/, accessed May 2015.

[8] B. Makoond, ZDLC Overview, http://0deviation.com/zdlc-platform/overview/, accessed May 2014.

[9] Deeptimahanti, Deva Kumar and Ratna Sanyal. "Semi-automatic generation of UML models from natural language requirements." Proceedings of the 4th India Software Engineering Conference. ACM, pp 165-174, 2011.

[10] Ambriola, V., & Gervasi, V. (1997, November). "Processing natural language requirements". In Automated Software Engineering, 1997. Proceedings., 12th IEEE International Conference (pp. 36-45). IEEE, 1997.

[11] Goldin, L., Berry, D.M.: "AbstFinder, a prototype natural language text abstraction finder for use in requirements elicitation". *Automated Software Eng*., pp.375–412, 1997.

[12] Kof, Leonid, "Natural language processing: mature enough for requirements documents analysis?" *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, pp 91-102, 2005.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

An Optimization Method for Embarrassingly Parallel under MIC Architecture

Yunchun Li and Xiduo Tian
Sino-German Joint Software Institute
School of Computer Science and Engineering, Beihang University
Beijing, China
lych@buaa.edu.cn, tianxiduo@gmail.com

***Abstract*—Nowadays, heterogeneous architecture of CPU plus accelerator has become a mainstream in supercomputing. Intel lauched its Xeon Phi coprocessor in this context. It uses Intel's many-core architecture, which greatly improves the single node parallelism. This paper studies the optimization of embarrassingly parallel programs under Intel MIC architecture, to maximize the utilization of CPU and Phi processor, and reduce the running time of parallel programs, by combining the computing power of CPU and Phi. This so-called embarrassingly parallel program often have doall main loops, that is, there are no dependencies between iterations, so they can be fully parallelized. This doall loop exists in many typical parallel programs. We come up with a loop allocation method for doall loops under this CPU/MIC architecture, to satisfy the above performance objectives.**

Keywords-exascale; many-core; embarrassingly parallel; loop allocation; performance tuning; Intel Xeon Phi

I. INTRODUCTION

At present, supercomputing has entered the era of petascale, the computing power of supercomputers is likely to reach exascale in the next few years [1]. To improve single-node parallelism, the supercomputers today tend to use the CPU+X way, where X means GPGPU or MIC. The latest Top500 supercomputer Tianhe-2 and Stampede both use the Intel Xeon Phi coprocessor to achieve node-level acceleration. In this context, the research of performance tuning of parallel programs under MIC architecture becomes very meaningful.

MIC combines the features of traditional general-purpose CPU and dedicated accelerator [2][3]. A typical Phi processor has 60 cores, each of which support 4 hardware threads, thus it can support a total of 240 hardware threads, with a parallel degree of 240. With MIC, parallel programs usually run in an offload mode, that is, CPU is responsible for the logical calculation, when it encounters a parallel region, it offloads the code to execute on Phi processor. Phi processor also supports the MIC native mode, when CPU doesn't participate in the calculation at all [4][5].

In MIC architecture, most parallel programs can improve their performance through the CPU/MIC offload way. But the question is, how to maximize the utilization of CPU and MIC, and how to achieve load balancing between CPU and MIC. Of all parallel programs, there is a class of ideal parallel programs, called embarrassingly parallel, they use doall loop as their main loop, with no dependencies between iterations, so they can be fully parallelized. For this kind of parallel programs, we can split the main loop into two parts, and allocate the iterations properly between CPU and Phi processor to achieve the load balancing between CPU and Phi.

The experiment of this paper is based on the well-known NAS Parallel Benchmark [6] by NASA. Wherein, EP program is the so-called embarrassingly parallel program mentioned before. By rewriting the EP program, we propose an optimization method for doall loops under MIC architecture. By making comparative experiments with the rewritten version and the original version of EP program on CPU and Phi processor, we verify the feasibility and effectiveness of this approach. The rest of the paper is organized as follows: Section II shows the related research, pointing out the shortcomings of the existing optimization method under MIC architecture. Section III gives a detailed description of the optimization method of embarrassingly parallel program under MIC architecture. Section IV shows the experiment results of NPB EP, and makes a brief analysis of the results, as a verification and discussion of the method proposed in Section III. As a conclusion, Section V summarizes the significance and deficiencies of the past work, and lays the foundation of the future work.

II. RELATED WORK

The related work of parallel performance tuning under MIC architecture can be classified into the following categories: improve the performance of existing parallel programs by combining the computing power of CPU and Phi processor; distribute tasks between CPU and Phi processor using a variety of task assignment algorithms to satisfy certain performance requirements; migrate existing parallel benchmarks to the Phi processor.

[7] gives the results of NPB programs running in native mode on CPU and Phi processor respectively, and compares the running time difference between CPU and Phi. This approach simply recompiles the original code under MIC architecture, and runs it natively on Phi processor, thus Phi processor is responsible for all the workload, CPU and Phi are not combined to improve performance.

[8] introduces the classic loop allocation policy in parallel programs, including static allocation and dynamic allocation, which is also used in OpenMP loop allocation strategy. However, these loop allocation policies are limited to one single OpenMP environment. In MIC architecture, the loop allocation policies cannot be used directly, because there are two separate OpenMP environments, namely CPU and Phi. Therefore we have to reconsider how to achieve an efficient loop allocation between CPU and Phi processor

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.12

under this hybrid OpenMP environment. Hence, this paper proposes an adaptive loop allocation method under MIC architecture, which enables high degree of parallelism between CPU and Phi processor, and optimizes the loop execution under MIC architecture.

[9] is a previous work of our lab on NPB performance tuning under MIC architecture, which proposes a task assignment method for Phi processor, and uses NPB CG as an example. First, establish the task dependency graph by analyzing the existing parallel program. Then, add interested performance factors in the task dependency graph as weights of vertices, which means tasks, and edges, which means communications bewteen tasks. Finally, calculate the optimal task assignment graph between CPU and Phi processor, so that the performance is optimized under certain performance factor. This paper mainly uses a static task assignment method, it optimizes a certain performance factor, such as the memory usage of Phi processor, by manually modify the task distribution between CPU and Phi processor using the task assignemnt result above.

III. METHOD DESCRIPTION

Typically, there is a main loop in a parallel program responsible for most calculation of the program, with a typical structure of for loop.

In MIC architecture, to maximize the utilization of CPU and Phi processor, and reduce the execution time of parallel programs, CPU and Phi processor should be combined using offload mode, because either CPU native mode or MIC native mode causes the other to be in idle. In offload mode, by using offload signal statement, you can offload MIC code onto Phi processor and return immediately to execute the CPU code without waiting for the MIC code to be completed, then by using offload_wait statement, you can insert a synchronization point at the end of the CPU code and synchronize the CPU code and MIC code at this point. In this way, we achieve parallelism between CPU and Phi processor.

The loop allocation problem under MIC architecture is to decide how we split the main loop into two reasonable parts, and distribute them between CPU and Phi processor. The result of the loop split shoule be that CPU and Phi processor arrive at the sync point almost simutaneously, thus furthest minimizing the wait time of each other.

There are two loop allocation methods here, one can be called static way, the other can be called dynamic way, both of which are based on sampling the execution time of CPU code and MIC code respectively, thus detemining the computing power of CPU and Phi, and further determining how the iterations distribute between CPU and Phi processor. The static way determines the number of iterations on CPU and Phi respectively before running a parallel program in offload mode. The dynamic way adjusts the number of iterations on CPU and Phi adaptively when a parallel program is running until an optimized allocation is achieved.

The static way requires pre-execution of the program on CPU and Phi processor respectively, with timers inserted into the program at the beginning and end of the main loop.

On this basis, we can decide the average time of each loop iteration on CPU and Phi processor respectively, then we can further determine the loop allocation factor between CPU and Phi processor. The static way uses one-time allocation, so there needs only one single data transfer between CPU and Phi processor. Fig.1 gives a brief description of the static allocation.

```
//CPU Native
timer_start(1);
for (k = start; k < start + iterations; k++) {
   //iteration k
}
timer_stop(1);
t1 = timer_read(1);

//Phi Native
timer_start(2);
for (k = start; k < start + iterations; k++) {
   //iteration k
}
timer_stop(2);
t2 = timer_read(2);
factor =1/ (1 + t1 / t2);

//CPU offload
for (k = start; k < start + factor * iterations; k++) {
   //iteration k
}

//MIC offload
for (k = start + factor * iterations; k < start + iterations;
k++) {
   //iteration k
}
```

Figure 1. Static Loop Allocation

An ideal solution for dynamic allocation is to use the producer-consumer model, which consists of three parallel tasks, a producer reponsible for the loop allocation, two consumers namely CPU and Phi processor requesting iteration chunks from the producer. The producer produces two kinds of iteration chunks with different sizes based on the consuming ability of the two consumers. When either consumer completes their iteration chunk, it requests another one from the producer. The two consumers are independent of each other and no synchronization is needed. When the producer has given out all the iteration chunks, the whole task is completed. However, this ideal solution is hard to realize, because there is not a third party besides CPU and Phi processor to be a producer, as it is not possible for the CPU to be a producer while serving as a consumer.

For this reason, we propose an alternative solution. CPU first offloads the MIC iteration chunk to Phi processor and return immediately, then it tries to carry out several CPU iteration chunks until the MIC iteration chunk is completed.

As the single core computing power of MIC is weaker than that of CPU, the MIC iteration chunk executes longer than the CPU iteration chunk, so we can insert several CPU iteration chunks during the execution of a single MIC iteration chunk. The iteration chunk size of CPU and MIC is determined by the hardware parallelism of CPU and Phi processor, that is, the supported hardware thread number of CPU and MIC. The benefit of dividing iterations into chunks according to processor hardware parallelism is that we can make full use of the processor's computing power but without the overhead of task switching.

```
while(true) {
// main loop iteraion
#offload target(mic) signal

(MIC CODE)
timer_start(1);
#pragma omp parallel for
for (k = start1; k < start1 + mic_threads; k++) {
   //iteration k
}
timer_stop(1);
t1 = timer_read(1);

(CPU CODE)
timer_start(2);
#pragma omp parallel for
for( k = start2; k < start2 + cpu_threads * factor; k++) {
   //iteration k
}
timer_stop(2);
t2 = timer_read(2);

Method (1)
if (t1 > t2)
  factor += 1;

Method (2)
factor = t1 / t2;
#offload_wait wait
}
```

Figure 2. Dynamic Loop Allocation

In this way, we tranform the main loop into a few MIC iteration chunks and CPU iteration chunks. In the original main loop, we execute an iteration at a time; in the tranformed main loop, we execute a MIC iteration chunk bundled with a few CPU iteration chunks at a time. CPU and MIC run asynchronous, so they are almost parallel. In each iteration of the tranformed main loop, we increase the number of CPU iteration chunks gradually, until the execution time of a MIC iteration chunk is nearly equal to that of a certain number of CPU iteration chunks. The question is how we can decide the number of CPU iteration chunks bundled with a MIC iteration chunk. There two ways to choose from: one is that we linearly increase the iteration chunks of CPU by one chunk a time until the execution time of a MIC iteration chunk is nearly equal to that of CPU iteration chunks; the other is that we determine the number of CPU iteration chunks based on the ratio of the execution time of a MIC iteration chunk and a CPU iteration chunk. The latter is faster, but may be inaccurate due to the use of only one iteration. The former is slower, but more accurate due to the use of more than one iteration. The latter can be combined with the former to get better results, that is, we use the latter way first, and then we use the former way to adjust the result. Fig.2 gives a brief description of the dynamic allocation.

IV. Experiment Results and Discussion

The experiment of this paper is carried out on a workstation configured as follows: two Xeon E5-2609 CPUs, 2.4GHz clock speed, 32GB main memory, CentOS 6.4 operating system. The workstation has a Xeon Phi 5110P coprocessor connected via PCI-E bus, with an MPSS version of 2.1.6720-13. The EP program comes from the OpenMP version of NPB 3.3. The hardware parallelism degree of CPU is 8; the hardware parallelism degree of Phi processor is 240, but in offload mode, it reduces to 236 because in offload mode, there must be a core responsible for data transmission and communication between CPU and Phi processor, so the cores available for computing reduces to 59.

The experiment results are shown in Table 1 and Fig.3. The first two rows of the table shows the execution time of EP running in native mode on CPU and Phi processor respectively. Based on the first two rows of the result, the static allocation policy works out an optimal allocation factor of about 1:2 between CPU and Phi processor, i.e., 1/3 iteraions on CPU and 2/3 iterations on Phi. As a comparative experiment, we give the results of 1/2 CPU 1/2 Phi and 1/4 CPU 3/4 Phi, in which CPU is overloaded and underloaded respectively.

As can be seen from the table, the execution time of EP using dynamic loop allocation policy is very close to that using static allocation policy. However, because the static way only transfers data between CPU and Phi once, the execution time is less compared to dynamic way. One thing to note is that the scale factor method is actually combined with linear increase method after it works out a proper scale factor.

In our experiment, we see that the loop allocation factor between CPU and Phi processor becomes stable after a certain number of iterations. The loop allocation factor between CPU and Phi is 17, which means we can bundle 17 CPU iteration chunks with 1 MIC iteration chunk. We got this allocation factor by repeatly printing the variable that represents the allocation factor at each main loop iteration. We can compare this factor with the factor obtained by the static way. When the ratio of CPU iteration chunks and MIC iteration chunk is 17, the actual loop allocation factor is $236*1/8*17 \approx 1.74/1$, that is, CPU accounting for $1/(1+1.74)=0.37 \approx 1/3$, and MIC accounting for $1.74/(1+1.74)=0.63 \approx 2/3$.

TABLE I. THE EXECUTION TIME OF NPB EP UNDER DIFFERENT LOOP ALLOCATION POLICIES (UNIT: SECONDS)

Time(EP)		Class S	Class W	Class A	Class B	Class C	Class D
CPU Native		0.08	0.16	1.28	5.09	20.26	323.29
Phi Native		0.08	0.11	0.74	3.00	11.19	185.36
Static	1/2CPU+ 1/2Phi	1.87	1.93	2.17	3.39	10.89	169.91
	1/3CPU+ 2/3Phi	1.87	1.92	2.29	3.64	8.94	115.54
	1/4CPU+ 3/4Phi	1.92	1.93	2.38	3.85	10.02	130.10
Dynamic	CPU+Phi (linear increase ethod)	1.93	1.98	2.44	3.86	9.67	128.22
	CPU+Phi (scale factor method)	1.96	1.98	2.42	3.84	9.71	128.48

To further understand the execution behavior of EP under different loop allocation policies, we made timeline analysis of different policies using Intel VTune Amplifier. We choose EP of CLASS C in this step. In this way, we have a relatively large problem size, so that we can see the acceleration effect of MIC, and we avoid the use of CLASS D which is a very time-consuming problem size. The experiment results are shown in Fig.4.

In Fig.4, the horizontal axis represents execution time, and the vertical axis represents EP using different loop allocation policies. In this Figure, all results come in pairs except for CPU Native and Phi Native on the top. As we can see from the figure, for static allocation of 1/3 CPU 2/3 Phi, the overall execution time is lowest; for static allocation of 1/2 CPU 1/2 Phi, CPU execution time is relatively long, and Phi execution time is relatively short due to load imbalance; similarly, for static allocation of 1/4 CPU 3/4 Phi, CPU execution time is relatively short, and Phi execution time is relatively long due to load imbalance. The load imbalance shown in VTune may be caused by VTune sampling overhead.

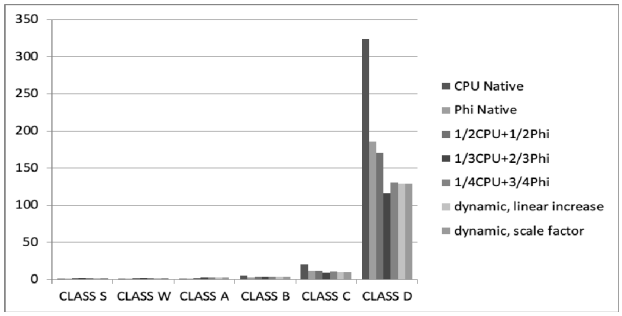


Figure 3. The Execution Time of NPB EP under Different Loop Allocation Policies

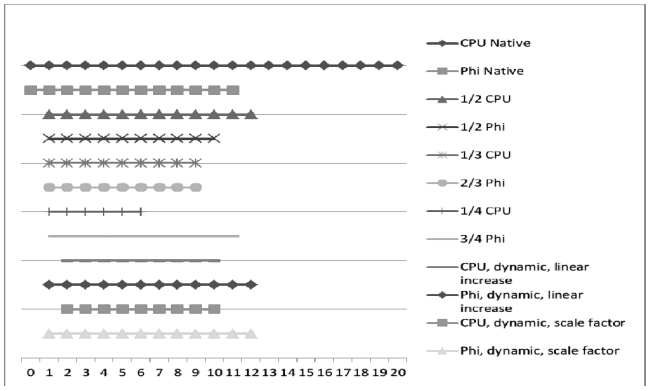


Figure 4. Summary of VTune Analysis Results

V. CONCLUSION

Through the experiments above, we show the effectiveness of the two loop allocation method, verify that they can significantly reduce the execution time and improve execution efficiency of parallel programs under MIC architecture. The deficiency of static loop allocation policy is that it needs pre-execution of the program on CPU and Phi processor, and the shortage of dynamic loop allocation is that it causes extra time overhead due to the frequent communication between CPU and Phi processor. We should make the right choice according to specific needs.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61133004) and National Hi-tech R&D program of China (863 program)(Grant No. 2015AA01A301).

REFERENCES

[1] Kogge P, Bergman K, Borkar S, et al. Exascale computing study: Technology challenges in achieving exascale systems[J]. 2008.

[2] Borkar S. Thousand core chips: a technology perspective[C]//Proceedings of the 44th annual Design Automation Conference. ACM, 2007: 746-749.

[3] Duran A, Klemm M. The Intel® many integrated core architecture[C]//High Performance Computing and Simulation (HPCS), 2012 International Conference on. IEEE, 2012: 365-366.

[4] Cramer T, Schmidl D, Klemm M, et al. OpenMP Programming on Intel R Xeon Phi TM Coprocessors: An Early Performance Comparison[J]. 2012.

[5] Jeffers J, Reinders J. Intel Xeon Phi coprocessor high-performance programming[M]. Newnes, 2013.

[6] Bailey D, Browning D, Carter R, et al. The nas parallel benchmarks, 1994[J]. NASA Ames Research Center: Moffett Field, CA, 1994.

[7] Ramachandran A, Vienne J, Van Der Wijngaart R, et al. Performance evaluation of NAS parallel benchmarks on Intel Xeon Phi[C]//Parallel Processing (ICPP), 2013 42nd International Conference on. IEEE, 2013: 736-743.

[8] Hurson A R, Lim J T, Kavi K M, et al. Parallelization of DOALL and DOACROSS loops—A survey[J]. Advances in computers, 1997, 45: 53-103.

[9] Li Y, Zhang T. A Task Assignment Method for Phi Structure[C]//Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2014 13th International Symposium on. IEEE, 2014: 38-41.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Metadata Namespace Management of Distributed File System

Baoshan Luo [1]
[1] School of Computer Science, Wuhan Vocational College of Software and Engineering
Wuhan, China
bsluo@163.com

Xinyan Zhang [1]
[1] School of Computer Science, Huazhong University of Science and Technology
Wuhan, China
zhangxinyanhust@gmail.com

Zhipeng Tan [2]
[2] School of Computer Science, Huazhong University of Science and Technology
Wuhan, China
zhipengtan@163.com

***Abstract*—With the rapid development of internet, the distributed file system is gradually facing problem of supporting concurrent 10K and 100K servers operations, so it now becomes an important issue that how to improve performance of the metadata server in high concurrent reading /writing environments. This paper designs and implements the distributed file system named RaccoonFS, researches and solves the problem of low response speed of namespace management brought by the lock competition in distributed file system, and it proposes a new metadata namespace management method that realized the technology of Copy-On-Write and Multi-Version Concurrency Control in B+ tree. Through the evaluation, the namespace management that combined COW and MVCC in B+ tree greatly improved the read and write performance.**

Key words: distributed file system; metadata management; copy on write; multi version concurrency control

1. INTRODUCTION

With the rapid development of internet, the data produced every day is to be calculated by PB levels and much of them should be persistently stored in hard disk. Traditional file system like ext2, ext3 and net file system (NFS) [1] cannot satisfy our needs, the distributed file system appears and becomes more and more popular. The typical distributed file system like HDFS [2] has been a basic platform for industry and academia. Distributed file system is made up of metadata server, data server and client, the metadata server is the brain of all nodes, how to enhance metadata management is important for improving the performance of the cluster.

People have made detailed study on data access and metadata access in distributed file system. When executing metadata access operations, the data package exchanged with the metadata server is relatively small, it is compute-intensive, the main consumption is server's computing resources, and users have a higher real-time requirements on this part; when executing the data access operations, users usually hope the system has high throughput, the requirement to response time is not higher than that of metadata access, it is I/O intensive and it mainly consumes the network bandwidth of servers. As it has the big difference between these two access operations, people then decouple them and put forward the metadata server and data server. Considering the data consistency, metadata is centrally located in a single machine that has strong service capacity, data servers can be deployed in multiple machines and they centrally register themselves to the metadata server. Such systems have strong scalability, like GFS [3] and HDFS, etc.

With the growth of data, capacity of the corresponding metadata is growing rapidly, and the single metadata server goes through the bottleneck as it limits the capacity of storage and computation. Multiple metadata servers are then put forward, they provide service as a whole and coordinate inner to finish tasks, typical systems are CEPH [4], Capella, etc.

A new system appears as time goes by, it is based on P2P technology and it organizes those unrelated machines to provide external services. The information about metadata exists in each machine of current system, so this system don't have the single point of failure, but the implementation of consistency and reliability are difficult for it, such as GlusterFS, OceanStore [5],Granary [6] etc..

This paper studies the key issues which affect the performance of metadata management in distributed file system, and investigates some existing metadata management methods in distributed file system. On the basis of the existing solutions, we combine some solutions in similar scene of other systems, put forward a solution supporting high concurrent read and write operations, design and implement a distributed file system named RaccoonFS.

2. RELATED WORK

This paper mainly focuses on the namespace management of RaccoonFS and studies the deficiency of traditional method of namespace management under high concurrency on read and write operations, then puts forward a new solution. The metadata server is mainly used to storage system files, namespace information of directory and replica information, it usually contains three information: mapping from file name to file ID, mapping from file ID to block ID, mapping form block ID to data servers (replica information). Namespace management is responsible for the mapping from file name to file ID and mapping from file ID to block ID.

The namespace management mainly contains three methods: centralized namespace management, distributed namespace management and no central namespace management. In the early emergency of distributed file system, the metadata server and data server are sharing a machine, this makes system appear bottlenecks. For realizing high scalability and high performance of system, people found that store metadata and data separately can reach better performance, then they put forward a centralized namespace management scheme. It contains three parts: multiple clients, data servers and single metadata server. It deploys all the metadata in one

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.13

metadata server. Unlike conventional storage access system, its metadata access and data access are not in one machine, it reduces load of metadata server, and it makes control information and data information concurrent, which greatly improves response time and throughput of system, but it has problem of performance bottleneck and single point of failure. The typical systems are HDFS, GFS, KFS, etc.

Distributed namespace management stores metadata information in multiple servers so as to enhance storage capacity of the entire cluster. The typical system is CEPH [8]. It resolves performance bottleneck and single point of failure in centralized namespace management method, however, it brings problem of too large performance overload and consistency problems.

No central namespace management is put forward with the development of P2P [9] technology, it is usually achieved by the consistent hashing. It has no central node and no single point of failure, it greatly improves reliability, stability, and expansibility of system, but it makes data consistency more complex and it lacks support of range quires, and it leads to the low operational efficiency of the directory. The typical system is GlusterFS.

The main work for metadata server is saving information of namespace and replica, most current distributed file systems take the centralized or distributed namespace management methods. After doing the research on the systems like HDFS, haystack, XFS, CEPH, it is common that massive users cooperate each other in main Web application runtime environment, although the fully distributed structure can support performance extension, but with the growth of the system scare, it is nearly impossible to ensure consistency. The centralized management is convenient to control cluster, it can be optimized after combining appropriate methods. We design a new management method that uses B+ tree to manage namespace, realize the COW and MVCC technology on B+ tree, separate read/write operations and make the write concurrency, it aims at improving the read/write performance and avoiding the performance overload brought by locking.

3. THE DESIGN AND IMPLEMENT OF RACCOONFS

RaccoonFS uses centralized metadata management, namespace management are important in design of centralized metadata management. With the rapid development of internet, more and more network services is facing problems of C10K [11] and C100K, metadata servers of distributed file system also have these problems. It tends to be inefficient and even is completely paralyzed, when metadata server dealing with tens of thousands of client connections, it is an important design element to improve the performance under high concurrent read and write operations.

3.1 Architecture of RaccoonFS

RaccoonFS adopts the centralized metadata management, it is typical tripartite structure, and the overall structure of RaccoonFS is shown in following figure 1.

As is shown in figure 1, metadata server is the master node that stores the node information of data servers and three-level metadata information which include mapping from file name to file id, file id to block id and block id to data servers, the metadata servers have multiple hot standbys so as to prevent a single point of failure. The master and slaves own the common virtual IP, they are managed by heartbeat. Once the master node happens to be crashed, the slave node can immediately get the virtual IP and provide services. Metadata information can be stored in the memory as its data volume is small, then regularly updating to the disk. Besides that, client don't need to visit metadata server every time as it can cache the data information, there is only control flow between client and metadata servers so as to exchange the location information of block, so the load is relatively small to the metadata servers.

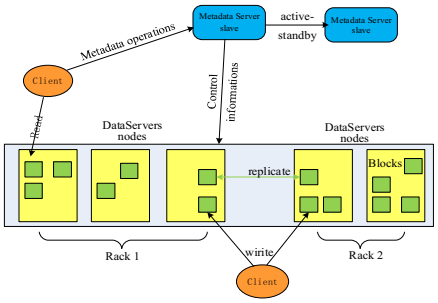


Figure 1 Architecture of the RaccoonFS

After dividing the big files, they are to be stored on data servers according to a certain load balancing strategy. The minimum storage unit of data server is the block (64M), each data has three standby (it can be configured). When executing the write operations, only all the data servers in pipeline succeed can it returns success so as to guarantee the strong consistency. The three standby can provide reading service when executing read operations.

Client is used for users, it provides users several simple interfaces for operations about reading, writing, deleting, etc. The distributed cluster is transparent to the user, after communicating with metadata server, client obtains the file allocation in the blocks and the mapping from the corresponding block to the servers IP address, then directly interacts with the data server and finishes the operations about reading, writing, deleting, file cutting, and merging.

3.2 Namespace management of RaccoonFS

RaccoonFS uses B+ tree [12] to manage namespace for making up the inadequate organization of HDFS directory-style. It realizes copy-on-write and multi-version concurrency control technology in the B+ tree, separates read and write operations, and it improves the performance of read and write and avoids the performance overhead brought by the lock.

3.2.1 Copy-on-write technology in RaccoonFS

As the read and write ratio of most applications is close to 10:1, read operation of copy-on-write technology [13] does not need to be locked, read and write operations do not affect each other, it greatly improves read efficiency and current server usually has 8 or 16 cores, so the COW technology is widely used in many companies.

We use tree managing metadata space, because it can avoid the operating expenses brought by uneven distribution directory, and the Copy-on-write technology is relatively easy to achieve in the tree structure. If we realize a B+ tree that supports copy-on-write technology, it can basically be used as internal data structures in most

management structures such as file and block management of RaccoonFS.

To realize the COW technology in B+ tree, the B+ tree is modified as follows:

1. It updates from top to down;
2. It removes links between leaf nodes in B+ tree;
3. It takes the reference count of a lazy ways, realizes the automatically release of the copied nodes.

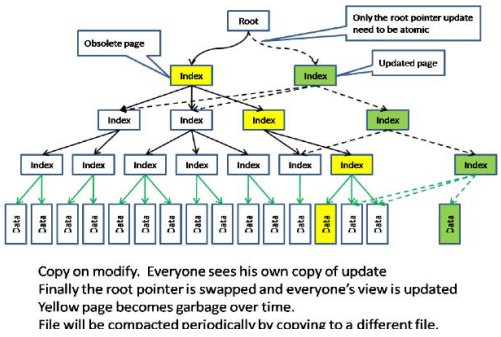


Figure 2 Copy-On-Write

As shown in figure 2, when B+ tree executes the updating operations, it will copy out all the nodes from the leave to root, perform modifications on the copied node, and the updating operation is submitted by atomic switching the root pointer. After adopting Copy-on-write technology, it will read old data if read operation occurs before the write operation takes effect, otherwise it will read the updated data and it does no need to be locked. The COW technology needs to use the reference count, when a node is not referenced, it means the reference count is released when it drops to 0, it greatly increases the complexity of the data structure. Besides, there is no interruption when doing the snapshot. If we need to take snapshot on a subtree of b-tree, we only need to add the reference count of root node to this subtree, subsequent read operations will read data produced by the snapshot.

RaccoonFS just uses the reference count for root node of the B+ tree, this reduces the complexity of implementation. It also uses a copy node queue to realize the automatic release of the copied nodes. Realization process is shown in figure 3.

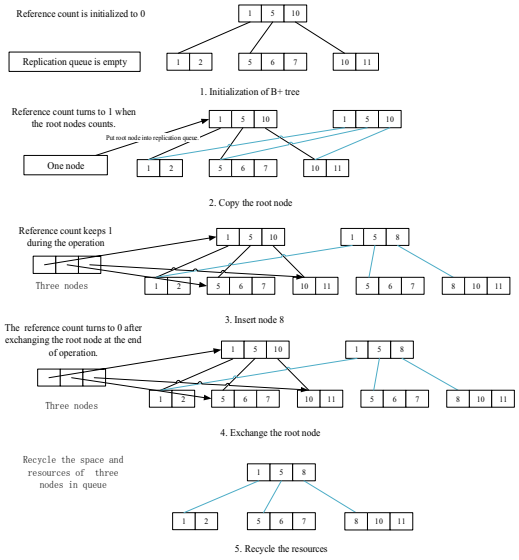


Figure 3 Realization of reference count

3.2.2 *Multi-version concurrency control technology in RaccoonFS*

Multi-version concurrency control technology [14] is an important concurrency control technology. Compared with the traditional two-phase locking technology [15], it ensures non-blocking between read and write operations, and keeps data consistency under multi-user concurrent operations.

We can conclude from the figure 4 that it will waste much space if the data's version item is overmuch and do not recycle for a long time. So we need recycle the related version items so as to reduce load of system when the read transactions do not need these related version items. If time-tamp of all read transactions in system is greater or equal to the time-tamp of version needed to be checked, then this version can be recycled.

RaccoonFS uses lock-free transaction mechanism based on MVCC technology. People call lock-based concurrency control mechanism pessimistic mechanism and MVCC is called optimistic mechanisms. Because locking mechanism is prophylactic, read and write operation will block each other, the concurrent performance wouldn't be good when the locking granularity is large and time is long. MVCC is a posteriori mechanism, read and write operation don't block each other as no lock existed, it tests whether there is a conflict until it submitted, which greatly improves concurrent performance. MVCC avoids large size and long locks, this make it better adapts to the scenarios that requires high concurrency and quick response to read operations.

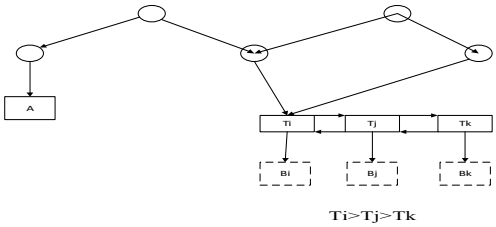


Figure 4 Multi version concurrency control

RaccoonFS uses version list to manage multiple versions of B+ tree [16], it consists of time-tamp, version pointer item and doubly linked list connected with version pointer item. The version pointer item is stored in ascending order according the time-tamp.

3.2.3 *Implementation of metadata server namespace management*

RaccoonFS uses B+ tree to manage directory structure and realizes COW and MVCC technology based on B+ tree, greatly improves performance of metadata server under high concurrency situations. In order to save some properties of block and mapping from file name to file id, file ID to block id. The node in B+ tree has below several properties:

META INTERNAL: it identifies the non-leaf nodes.

META DENTRY: it identifies nodes inner the leaf nodes, it saves mapping from file-name to file id.

META FATTR: it identifies nodes inner the leaf nodes which saves properties of file and mapping from file id to block id.

META BLOCKINFO: it identifies nodes inner the leaf nodes that saves information.

In B+ trees, only leaf nodes keep data, no-leaf nodes only keep index information so as to accelerate the search speed. There exists three types of leaf nodes in system: file directory node (*MetaDentry*), file attribute node (MetaFattr), data block node (MetaBlockInfo), its size can be compared among nodes and it decides its location in B+ tree. Through the compared function, the files or directories with the same parent node continuously arrange in the B+ tree, a block of the file in the file location arranged in a row. This structure can not only avoid the performance expenses brought by the uneven distribution of the traditional directory structure, but also can realize the operations like *ls*, *df*, *du* and other operations through the comparison function.

4. Evaluation and Analysis

4.1 Experiments platform

Our experiment cluster consists of 1 metadata server, 6 data servers and 120 clients. All nodes have eight 2.27GHz Intel Xeon CPU, 16GB main memory, the storage of each node contains a 500GB hard drive, all nodes are connected by 10GBps Infiniband network the operating system is RedHat Enterprise 5.4 with Kernel 2.6.18.

Both the stand-alone system and the distributed system have performance requirements. The following are the common performance indicators: system throughput (it means the amount of data that system can handle at one time, it can be balanced by the total number of data that system handle in a second), response delay (it means time that system used to finish one function), concurrent capacity (it means the capacity that system can simultaneously perform a certain function, it always use QPS (Query Per Second) to balance it). Unlike the IO-intensive data server (the main index is the throughput capability of system), the metadata server of distributed file system is compute-intensive, and it aims at the low latency and high response, so the higher the OPS is, the superior the performance is. The test mainly concerns about QPS.

Metadata server of distributed file system is typical of compute-intensive applications, its goal is low latency and high response speed, we mainly concern about OPS of RaccoonFS. To illustrate its metadata server have better service performance after RaccoonFS using new metadata namespace management mode, we choose three sets of test objects: namespace management methods of HDFS, B+ tree realized COW, B+ tree combined COW and MVCC, we separately do read, write, mixed read/write stress test on the three groups of test object, and test their QPS under the same pressure.

The operations of metadata server in distributed file system contain read and write operations, we test it from three aspects: response speed of concurrent reading test, response speed of concurrent writing test, response speed of mixed reading/writing test.

During the evaluation, we use one metadata server and 6 data servers, they are all located in the same local area network, we let 10 clients connect to the metadata server, the corresponding operations are carried out 1000 times, and record the required times, the OPS = (OpCount * number of clients) / (required time). And then we increase the number of clients according to the order of 20, 40, 60, 80, 100, repeat the corresponding operations for each clients, record the required time and the QPS = (OpCount * client numbers)/ (required time).

4.2 Response speed of reading test

RaccoonFS is similar to HDFS on read operations, we choose the typical *GetFileInfo* operation to do the comparison test.

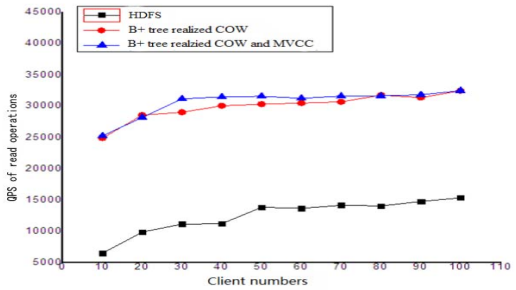


Figure 5 QPS of reading test

From the above figure, B+ tree realized COW and B+ tree combined COW and MVCC are roughly the same, the difference between them is the latter realize the MVCC, and MVCC is mainly used at optimizing write operation and improving writing concurrency, so the OPS of read operations are roughly the same, and they are both higher than that of HDFS, that is because the two former adopt multithread processing to improve response speed of read operations, and they both use C++ implementation and make many optimization in network libraries and reading processing.

4.3 Response speed of writing test

We choose the typical create operations to do the writing test. From the figure 6, it can be concluded that OPS of writing test in the B+ tree combined COW and MVCC is higher than that of B+ tree realized COW, it is because the former uses optimistic locking of MVCC on write operations, it makes write concurrency come true. The OPS of B+ tree that realized COW is not much higher than that of HDFS, it is because their mechanism are the same and both adopt the pessimistic locking, the writing concurrency cannot be come true and their operational processes are also similar.

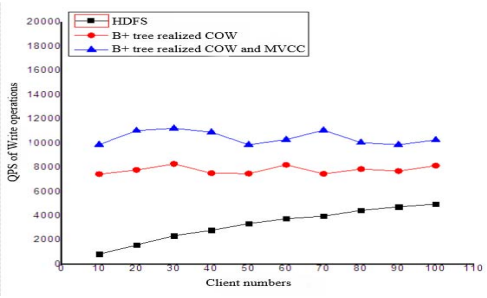


Figure 6 QPS of writing test

4.4 Response speed of mixed reading/writing test

We choose the operation of *GetFileInfo* in reading test and create in writing test to do the mixed test, the *GetFileInfo* operations are carried out 1000 times in 9 clients and the rest one client operates the create operations.

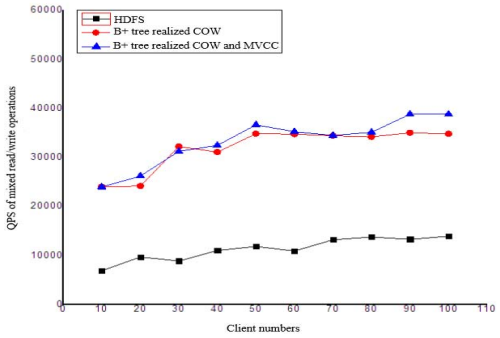


Figure 7 QPS of mixed test

Figure 7 shows the results of mixed reading/writing test, combined the results of 4.1 and 4.2, the QPS of mixed reading/writing test in B+ tree realized COW and in B+ tree combined COW and MVCC are basically equal to the sum of OPS in the separate reading and writing tests, but OPS of mixed test in HDFS is much lower than that sum of OPS in the separate reading and writing tests. It is because the former uses COW technology and comes true the separate of read and write operations, but read and write operations of HDFS are excluded and it needs an exclusive lock.

The result of the above three sets of test shows that the response speed of writing in B+ tree realized COW and MVCC is higher than that of B+ tree realized COW, the response speed of reading and writing in B+ tree realized COW is higher than that of HDFS, so the namespace management used B+ tree realized COW and MVCC can effectively improve the efficiency of the namespace management.

5. CONCLUSION

This paper analyzes the solutions about metadata namespace management in distributed file system, researches and resolves the problem of low response speed of distributed file system because of the lock competition, and realizes the RaccoonFS based on this project. RaccoonFS uses B+ tree method to manage the namespace and realizes the COW and MVCC technology on B+ tree, realizes the write concurrency and the separation of read and write operations, it greatly improves the performance of read and write operations and avoids the performance overload brought by lock. Through the three types of pressure test, its performance is superior to the traditional namespace management (directory management of HDFS).

REFERENCES

[1] Pawlowski B, Juszczak C, Staubach P, et al. NFS Version 3: Design and Implementation[C]//USENIX Summer. 1994: 137-152.

[2] Shvachko K V. HDFS Scalability: The limits to growth[J]. login, 2010, 35(2): 6-16.

[3] Ghemawat S, Gobioff H, Leung S T. The Google file system[C]//ACM SIGOPS operating systems review. ACM, 2003, 37(5): 29-43.

[4] Weil S A, Brandt S A, Miller E L, et al. Ceph: A scalable, high-performance distributed file system[C]//Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, 2006: 307-320.

[5] Kubiatowicz J, Bindel D, Chen Y, et al. Oceanstore: An architecture for global-scale persistent storage[J]. ACM Sigplan Notices, 2000, 35(11): 190-201.

[6] Zheng W, Hu J, Li M. Granary: Architecture of object oriented Internet storage service[C]//E-Commerce Technology for Dynamic E-Business, 2004. IEEE International Conference on. IEEE, 2004: 294-297.

[7] Varade M, Jethani V. Distributed Metadata Management Scheme in HDFS[J]. International Journal of Scientific and Research Publications, 2013, 3(5).

[8] Weil S A, Brandt S A, Miller E L, et al. Ceph: A scalable, high-performance distributed file system[C]//Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, 2006: 307-320.

[9] LAI H, WANG P. P2P Traffic Control Based on Ternary Content Addressable Memory[J]. Computer Engineering, 2010, 9: 043.

[10] DeCandia G, Hastorun D, Jampani M, et al. Dynamo: amazon's highly available key-value store[C]//ACM SIGOPS Operating Systems Review. ACM, 2007, 41(6): 205-220.

[11] Liu D, Deters R. The reverse C10K problem for server-side mashups[C]//Service-Oriented Computing–ICSOC 2008 Workshops. Springer Berlin Heidelberg, 2009: 166-177.

[12] Toptsis A A. B**-tree: a data organization method for high storage utilization[C]//Computing and Information, 1993. Proceedings ICCI'93., Fifth International Conference on. IEEE, 1993: 277-281.

[13] Braginsky A, Petrank E. A lock-free B+ tree[C]//Proceedings of the twenty-fourth annual ACM symposium on Parallelism in algorithms and architectures. ACM, 2012: 58-67.

[14] Tianhua L, Hongfeng Z, Guiran C, et al. The design and implementation of zero-copy for linux[C]//Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on. IEEE, 2008, 1: 121-126.

[15] Son S H, David R. Design and analysis of a secure two-phase locking protocol[C]//Computer Software and Applications Conference, 1994. COMPSAC 94. Proceedings., Eighteenth Annual International. IEEE, 1994: 374-379.

[16] Wu S, Jiang D, Ooi B C, et al. Efficient B-tree based indexing for cloud data processing[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1207-1218.

Research on Distributed Multimedia System In Universities Management Mode

LINGHU Xing-Rong
Guizhou Colloge of Finance and Economics
GuiYang ,China
lh@mail.gzife.edu.cn

***Abstract*—**This paper introduces the distributed multimedia system and its characteristics, discusses the application of distributed multimedia system in the field of teaching , And distributed multimedia system management mode in colleges and universities are discussed in this paper. The traditional multimedia equipment management in the real world problems is analyzed, and put forward some improvement measures on the basis of these problems, on this basis, the distributed multimedia equipment management in colleges and universities pattern carries on the preliminary exploration and research.

Keywords- Distributed multimedia; QoS management Modern teaching technology

I. I INTRODUCTION

The advent of the era of information, advanced technology and equipment are constantly applied to every corner of the society, the education sector is no exception, with the change of the modern education concept and the development of multimedia technology, also in changing the teaching way, colleges and universities, more and more investment in multimedia equipment, makes the teaching resources of colleges and universities are also improved. , however, the existing equipment management mode, the multimedia still exist factors hinder the normal development of multimedia teaching, how to better the application of multimedia equipment, improve its management mode, become the hot topic of colleges and universities[1].

In today's society has stepped into the information age, with the deepening of the reform of talent cultivation technology and the popularity of computer, the traditional teaching methods cannot meet the need of science and technology changes with each passing day. Represents the information revolution of the development of network technology and multimedia technology, promotes the modern education should to multidimensional, the direction of intelligent, personalized, wan, caused the change of human thinking and ways of learning. The teaching system based on network technology and multimedia technology is a research focus in the modern education technology, as well as the important direction in the field of application.

II. A DISTRIBUTED MULTIMEDIA SYSTEM

Distributed multimedia system is a multimedia technology, computer technology and network technology development of the product[2]. For distributed multimedia system are formally defined as follows: distributed multimedia system is an integration of the communications, computer information system, it to the synchronous information processing, management ,transmission, implementation has the service and quality assurance.

From the point of view of research and development, the distributed multimedia system has the following features:
1.Comprehensive. Usually, the information collection, storage, processing, transmission through different media. A single collection, storage and transmission media has its own theory and a series of special technology. And the multimedia integration refers to the different media, different types of information using the same interface, unified management. Here refers to unified management, said that they can be stored in a file, the information of different media can convert each other, in the form of system of different media information can automatically transform. This will greatly improve the efficiency of computer application, expand the application range.
2. Resources dispersion. This is a basic feature of distributed multimedia. It is different from the current multimedia systems, especially multimedia personal machine. It is all the resources of a centralized , the system is a single user. And what we call a distributed multimedia system resources dispersion refers to the system in various physical resources and logic resources, and on its function are geographically dispersed. Generally this system is based on client/server model. Using open mode, many of the nodes in the system of users to share resources on the server.
3. Run. Real-time multimedia, audio and video signals are related to time and continuous media. It's on the system real time request, such as the transfer film, a few pixels error innocuous, and suddenly waiting for error correction is irritating. The key problem is how to match the resources within the computer multimedia information, combination, form a whole.
4.Interactive operation. Refers to send in a distributed system, transmission and receive real time multimedia information using interactive mode of operation, namely, ready to carry on the processing of multimedia information, processing, modification, zoom in and back together.

III. THE STATUS OF MULTIMEDIA EQUIPMENT MANAGEMENT IN UNIVERSITIES

A. THE STATUS OF MULTIMEDIA EQUIPMENT MANAGEMENT IN UNIVERSITIES

1. Multimedia classroom management is not science

The multimedia classroom often belong to different department or departments, adopt different management pattern, causes the management efficiency is not high, due to the time of the purchase of multimedia equipment, model is

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.14

different, lead to the whole system of multimedia classroom management confusion, lack of systematic, is not conducive to effective management. Moreover, most multimedia classroom computer were independent of each other between, failed to form a joint network, also to the use of multimedia devices communicate with each other.

2. Lack of training for teachers.

In the multimedia equipment, the introduction of the lack of teachers' training. Teachers for the new multimedia devices are not familiar with, will not only lead to low efficiency of classroom teaching, possibly because the teacher on the operating error cause the damage of equipment, influence the normal teaching, therefore, it is necessary to the necessary technical training for teachers.

3. Equipment management personnel service consciousness is not strong.

Part of the management personnel set up the correct responsibility consciousness and consciousness, service consciousness is not strong, lack of patience, in the process of work for teachers and students gruff, rhetoric, easy to cause contradiction, influence the normal teaching consciousness.

4. Multimedia equipment purchase fund shortage.

At present, the colleges and universities needs to be faced funds tense situation, lead to many colleges and universities set up a multimedia classroom, although still cannot meet the demand of normal teaching. Colleges and universities for multimedia equipment investment funds is limited, even bought multimedia devices, but I did not put in enough money to regularly update and maintenance equipment and equipment utilization has great, maintenance is a big demand, funds and the contradiction between demand.

B. The improvement measures of existing multimedia devices use problems

1. Integration of multimedia equipment, set up a system of equipment.

Colleges and universities can be within the multimedia equipment shall be carried out in accordance with the type classification, as far as possible will be unified model of equipment placed in the same building, to improve the working efficiency of the daily equipment maintenance.

2.Establish rules and regulations, standardizing management.

Only truly established a standardized management system, to ensure the multimedia equipment management work systematic, improve the quality and efficiency of the management, to lay a solid foundation for multimedia equipment management work.

3.To strengthen the training of management personnel.

School can be a variety of ways, in does not affect the normal work under the premise of partial training management personnel, content mainly includes two aspects: one is through the study methods such as improving the level of business management, the practice experience of communication between internal management personnel, complement each other, make progress together. Secondly, improve the level of management thought, change ideas, establish service consciousness, to build a harmonious multimedia equipment management team, to create a good teaching environment for teachers and students.

4.Improve the efficiency of the use of multimedia equipment.

Only in the process of using modern technology and equipment to the real work, if not, then the advanced equipment is just a pile of waste, therefore, we need to arouse the enthusiasm of teachers use multimedia equipment,

IV. CONSTRUCTION OF NEW MODE OF MULTIMEDIA SYSTEM MANAGEMENT

The new model of multimedia equipment management is essentially a kind of new ideas. This model is mainly based on the concept of system, on the basis of open management, equipment management in colleges and universities as a unified system, the management of multimedia equipment work conducted by the school authorities unified scheduling and arrangement. In the process of management, between different departments to cooperate with each other, people-oriented, will better serve the work on the basis of the teachers and students, so as to eventually achieve the efficient use of multimedia equipment in colleges and universities and maximize the benefit of management.

Introducing teaching system based on distributed multimedia technology in classroom teaching, can accelerate the learning process of perception, promote deepening understanding, deepen understanding, enhance memory and improve the application ability, it pay attention to exert students' initiative, to improve students' cultural and psychological quality, cultivate students' innovative ability. Students choose to study, on the other hand, the content, time, difficulty and progress, have a greater degree of freedom. Teaching system based on distributed multimedia technology to design to satisfy the needs of the development of modern society, economy, science and technology, aimed at cultivating modern talents ` new model of modern education[3]. It covers the ideas of education modernization, modernization of education content, education facilities modernization and modernization of education management, it is a system

A. Distributed multimedia system model

<table>
<tr><td>Distributed multimedia application layer</td><td rowspan="5">System management</td></tr>
<tr><td>Distributed multimedia applications to support the platform layer</td></tr>
<tr><td>Kanadeai layer</td></tr>
<tr><td>Integrated transfer protocol layer</td></tr>
<tr><td>Many services the network layer</td></tr>
</table>

Figure1 Distributed multamedia system model

According to the domestic and foreign research results, based on [5], we proposed an improved distributed multimedia system model, as shown in figure 1. In the distributed multimedia application layer distributed multimedia application supporting platform based on the QoS layer of local resource management kanadeai integrated transfer protocol layer multimedia computer system hardware platform integrated services network layer management group communication FDT figure 1 improved distributed multimedia system model.

<table>
<tr><td colspan="2">Distributed multimedia application layer</td><td rowspan="5">System management</td></tr>
<tr><td colspan="2">Distributed multimedia applications to support the platform layer</td></tr>
<tr><td rowspan="2">Based on the QoS of local resource management</td><td>Kanadeai layer</td></tr>
<tr><td>Integrated transfer protocol layer</td></tr>
<tr><td>Multimedia computer hardware platform</td><td>Integrated services network layer</td></tr>
</table>

Figure 2 Improved distributed multimedia system model

B. Logical structure and the working process of the QoS management model

The Service quality(QoS) refers to the service performance of the cluster effect, it decided to treat service user satisfaction. A distributed multimedia system to deal with the information can be divided into two types of static media and continuous media. Continuous media data recording, access, transfer and playback process has strong real-time and isochronism, can produce a great amount of data in real time. Before transmitting multimedia data on the Internet, therefore, the key technology research is typically the source of compression, and broadcast after decompression at the destination, in the processing of continuous media data, not only need to maintain continuity within the same media time and usually need to maintain the synchronization between different media relations, for example, video and audio in the film "in synch" relationship between the mark. Therefore, distributed multimedia system management faces new challenges mainly comes from the continuous media data. According to the distributed multimedia system model shown in figure 2, we propose a QoS management model. In the QoS management model, the system of each layer and multiple services in the network exchange node configuration of a QoS Manager QoSM (QoS Manager). Below, we described the working process of the model based on Client/Server paradigm.

C.Three phase QoS negotiation protocol

Distributed multimedia applications users want a distributed multimedia system provides a certain degree of assurance, therefore, before the use of the service, users should request notification system, its necessary to negotiate, in order to agree on the parameter value, make the agreement of these parameter values mutually by user and the system of "contract".

The server before the services provided to clients, both sides need to first determine the QoS through three stages of QoS negotiation agreement.The first,the client by the forward QoS negotiation and resource reservation phase.Then the server initiates to resource reservation after relaxation stage.The last the client initiates the resource allocation phase/reservation cancelled.

QoS negotiation agreement after the end of three phase, "the contract" way for QoS[6]. In the process of negotiations, only consultation between the layers and peer consultation service network layer is more direct, and top peer consultation is based on low consultations and peer consultation among layers.

D.Dynamic QoS management process

After the three stages of QoS negotiation protocol, the server will start to customers according to the consensus of QoS.[7] During the period of service, the client and server side each QoSM in accordance with the agreed QoS "contract" to finalize the design of traffic, to meet the QoS requirements. Client system and network switching nodes in each QoSM monitoring end system and actual network switching nodes in each layer QoS level. If due to load changes in network and end system causes such as QoS degradation, the QoS management mechanism for dynamic adjustment. Below, we introduced the dynamic management process of QoS model.

1. Exchange network nodes by fine-grained QoS control

If found a network switching nodes QoS degradation, the node QoSM first by resource use this node optimization scheduling, striving to restore to the original value of QoS level, called the local regulation. If the local regulation, the local QoSM upstream (close to the server side) and its direct negotiations QoSM, ask whether it can reduce the input pressure of this node, so that the QoSM can through the release of other communication source used to restore some of the resources of QoS. If failed, and its direct downstream (close to the client side) QoSM negotiations, asking them whether to compensate, so as to recover in the downstream of QoS. We call this process chain two-way adjustment. If direct QoSM upstream and downstream nodes are powerless, then the request further spread to more upstream and downstream nodes, until she reached the server side QoSM QoSM or client side. If QoS adjustment request to reach the server, the server side QoSM try to adjust the QoS; If can't be reinstated, the QoS request over to the customer to QoSM adjustment. QoS control request arrives at the client side, the network exchange nodes initiate the QoS of QoS

adjustments of the adjustment translates as customer initiates.

2.The client initiates the coarse-grained QoS control

QoS demoted if customer QoSM discovery, the QoSM first internal resources of the system to adjust to end, for example, take a look at whether caused by application by the user to open too many system resources nervous, if it is, is advising clients to turn off some secondary application to make the necessary resources, to restore the QoS. If failed, then launch the end-to-end QoS negotiation, process such as the three phase QoS negotiation protocol.

3.QoS control on the solution of the "conflict"

If two or more networks due to switching nodes also found an application QoS degradation, thus initiated the application QoS control at the same time, the upstream nodes "absorb" downstream QoS request adjustment. If due to network nodes and the customer exchange also found an application QoS degradation, thus initiated the application QoS control at the same time, the network switching nodes "absorb" customer QoS control request, namely only when fine-grained adjustment is invalid, just a coarse-grained regulation.

V. CONCLUSION

University of multimedia equipment in teaching of colleges and universities has made a significant contribution, in the future the development of education, the scientific and multimedia equipment management model has been more and more significant impact on education of colleges and universities. With the development of modern science and education technology, the reform of the management model has become an inevitable trend. Therefore, we should actively learn from the college of management experience, carefully summarize and explore the suitable teaching actual need, is advantageous to the utilization of teaching resources, convenient and mutual learning between teachers and students of management mode, make the advantage of multimedia equipment better, more comprehensive, create a better learning environment. In a word, using scientific management mode, formation of less investment, good circulation, vomit can really promote the development of multimedia teaching quickly.

QoS management is to design a distributed multimedia system must solve one of the major problems. Due to the different QoS requirements of distributed multimedia applications, thus to further increase the difficulty that the design of QoS management mechanism[8].This paper proposes a QoS management model for , trying to solve the problem.

REFERENCES

[1] *Xie Jintao. Discuss about the problems existing in the construction and management of multimedia classrooms and countermeasure field. Network wealth, 2009 (8) : 37, 38.*

[2] *Shen Hai . The application of distributed multimedia system in teaching . Journal of shenyang normal university (natural science edition)[J], 2003, 22(1) : 34-38*

[3] *Zhang Qian . Based on the research of distributed multimedia teaching system. Nanjing university of technology, 2005*

[4] *Xing-wei wang, ying-hui zhang, liu jiren, huatian li. A QoS management model in distributed multimedia system [J]. Small microcomputer system, 1997, 19 (2) : 19-24.*

[5] *BHATTI SALEEM N. et al, Enabling QoS adaptation decisions for Internet applications[J].Computer Networks, 1999,31(5):669- 692..*

[6] *Yeung hok-leung, multimedia technology and application of computer, electronic industry press, Beijing, 1996*

[7] *Ran Shu-Ping.A model for Web service discovery with QoS[J].ACM SIGeeom Exchange. 2003.4(1):1-10*

[8] *Fang-xiong xiao, zhi-qiu huang, zi-ning cao, etc. The Web service composition function and the QoS of the formal modeling and analysis [J]. Journal of software, 2011, (22)2698- 2715 .*

Research on Petri Nets Parallel Algorithm Based on Multi-core PC

Zhi Zhong, Wenjing LI
School of Logistics Management and Engineering
Guangxi Teachers Education University
Nanning, China
e-mail: zhong8662@126.com

Yijuan SU, Ze-yu Tang
College of Computer and Information Engineering
Guangxi Teachers Education University
Nanning, China

***Abstract*—In order to make the concurrency, synchronism function of Petri nets system capable of parallel control and Simulation Implementation, proposed Petri network switching system, function partition method based on Multi-core PC. Firstly, according to the Petri nets system of parallel process and its principle, analysis of advanced Petri nets and P/T network algebra model and its inner mechanism, Given the process and theory verification of advanced Petri nets into P/T nets; Based on the network model Formalization and Color Petri nets correlation matrix pretreatment, proposed colored Petri nets into P/T nets algorithm. Then, P/T nets function division according to invariable Place technique, classification into subnets having different functions (process), analysis and expansion of the P/T nets system process conditions, and gives examples of authentication, Get P/T nets functional partitioning algorithm based on non-negative invariable Place; On this basis, Research process concurrency, synchronization parallel with implementation, put forward Petri nets parallel algorithm based on Multi-core PC, Given Petri nets parallel algorithms and application examples In the environment of Multi-core PC. Experimental results show that, Petri nets parallel algorithms based on Multi-core PC let Petri nets system to better reflect the actual running, and Is an effective method to achieve Petri net system parallel control and simulation run.**

Keywords- Multi-core PC; Petri nets conversion; Invariable Place; functional classification; Parallel analysis; parallel algorithm

I. INTRODUCTION

Petri nets parallelization process consists of three steps, the first is to extract the P / T net structure model, the second is the P/T nets invariants Place solving and sub netting (process), the third is for the subnet (process), and inter-process parallelization analysis and programming. Which extract P/T nets structure model is due to the wide variety of Petri nets model. But what kind of Petri net model of the system parallelism or functional division most convenient, the ideal, the issues that must be considered. Due to Petri nets too simple, can not effectively simulate the behavior of the system when transitions occur in the actual system running, transitions in the amount of state Place logo; Colored Petri Nets, timed Petri nets and some complex nets system extended on the basis of the P / T nets, due to contains redundant information in the functional division process variables, color, etc, and Not the ideal model of Petri nets parallelization. P/T nets model is to increase the capacity functions of the Place and weight function of directed edge set based on prototype Petri nets. Its can vividly describe the transitions in the system's structure, behavior, state library logo amount. Putting P/T nets model as parallelization object, Is an ideal choice. To this end, extract P/T nets structure model is about high-level Petri net system automatically translate into a P/T net. P/T nets invariants place solving and subnet are based on P/T nets invariants place defined and homogeneous linear equations, achieve the solution value of P/T nets system functions and classification according to the solution value. To parallel analysis subnet (process) and relative-process also is analysis concurrency, parallelism internal process and between relative-process, to solve Petri nets parallel mechanism. Nowadays, at home and abroad in the Petri net system running and parallel, literature [3] scan each transition about the nets system, used to examine transform trigger, proposed P/T nets centralization; literature [4] through the process to achieve each place and transition to make the Petri nets system distributed execution and proposed decentralized method of Colored Petri Nets. Centralized approach can not maintain the parallelism of the nets model; although Decentralized approach to ensure the nets system parallelism, But, when the Petri Nets scale becomes large and number of elements of the color set is large, the efficiency becomes very low. Literature [2] proposed invariants Place technology, by positive invariant library to build division process conditions. These studies exist three deficiencies: (1) The extract P/T nets structure model can not achieve automatically converted; (2) The empty place classification process conditions are not given subnet; (3) Formal representation of the building process, the method did not give two subnets exist a shared place. To this end, we research of Petri nets parallel algorithms In the Multi-core PC environment, and Propose solutions to the problems about Petri nets with concurrency, mutual exclusion, synchronization function parallel control and simulation run.

*Supported by the National Natural Science Foundation of China (61163012) ;Guangxi Natural Science Foundation (2014jjAA70175); the University Scientific Research Project of Guangxi(2013YB147)

II. EXTRACT THE P / T NETS STRUCTURE MODEL

Petri static nets model can not be executed, when simulate and run the nets model, need to generate parallel code or serial code. Petri nets parallelization process includes three steps, the first is to extract P/T nets structure model[5];

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.15

the second is solving the invariable place of P/T nets, and classification subnet (process); the third is parallelization analysis and programming the subnet (process) and relate-process. Therefore, the essence of extracting P/T nets structure model which is automatically converted the advance Petri nets into the P/T nets. So, the first research step is extract P/T nets structure model of the parallel process of Petri nets, which is to analysis between advance Petri nets and P/T nets internal mechanism. The next step is proposed the advance Petri nets automatically converted to P/T nets algorithm.

III. PLACE INVARIABLE PARTITIONING ALGORITHM

Definition 1[4] Let N= (P,T;F) is a nets, $|P|=m$, $|T|=n$, D is a associated matrix of N. If exist a non-trivial m-dimensional non-negative integer vector X satisfies $DX=0$, then X is an invariant place of the N net.

Definition 2[4] Take X is an invariant place of net N=(P,T;F),then $\|X\|=\{p_i \in P|X(i)>0\}$ is invariable place of branches set.

Theorem 1[6] Hypothes is of Γ_P is Petri nets N= (P, T; F, K, W, M) is invariable place branches sets, elements of Γ_P corresponding to the subnet are to content the following conditions: These elements corresponding to the subnet (process) corresponds to a N nets of a possible functional division, To classify the results get a group of independent function of the Petri nets process.

$$\forall p \in \{pi \in P|X(i)>0\}, \forall p \in (t,t') \in p\bullet \cap \bullet p \tag{1}$$

$$W(p,t) = W(t',p) = 1 \wedge \sum_{p \in \|X\|} M(p) \geq 1 \tag{2}$$

$$\bigcup_{i=1}^{|\Gamma_P|} \left(\bigcup_{j \in \|s_i\|} (p_j^{\bullet} \cup {}^{\bullet} p_j) \right) = T \tag{3}$$

$$\forall s1, s2 \in \Gamma p, \|s1\| \cap \|s2\| = \varnothing \tag{4}$$

Theorem 2[6] if the elements Petri nets invariants place branch set Γ p of N=(P,T;F,K,W,M) meet with the following conditions:

$$\forall p \in \{p_i \in P | X(i)>0\}, \forall p \in (t,t') \in p^{\bullet} \cap {}^{\bullet} p \tag{5}$$

$$(W(p,t) = W(t',p) = 1) \wedge \left(\sum_{p \in \|X\|} M(p) \geq 1 \vee \left(\sum_{p \in \|X\|} M(p) = 0 \wedge (M({}^{\bullet}(p^{\bullet})) \geq 1 \vee M({}^{\bullet}({}^{\bullet} p)) \geq 1) \right) \right) \tag{6}$$

$$\bigcup_{i=1}^{|\Gamma_P|} \left(\bigcup_{j \in \|s_i\|} (p_j^{\bullet} \cup {}^{\bullet} p_j) \right) = T \tag{7}$$

$$\forall s_1, s_2 \in \Gamma_p, \|s_1\| \cap \|s_2\| = \varnothing \tag{8}$$

Based on the above given Petri nets functional division, our place has been based on the place invariants of Petri nets function decomposition algorithm. The steps are as follows:

(1) Non- P/T nets of Petri nets models converted to P/T model, and order place to the transitions or transitions to the place have directed arc, weights for 1, there is no directed arc, weights were taken 0;

(2) Input Petri nets model of the output matrix $D^+=[d_{ij}^+]$, input matrix $D^-=[d_{ij}^-]$ and the initial identification $M=(M(p_1),M(p_2),...,M(p_m))$ initialized data etc;

(3) Through the formula $D = D^+ - D^-$ solve the associated matrix D ;

(4) Solve the homogeneous linear equations $DX=0$ to get the Petri nets place invariants collection;

(5)On place invariants in each element of the collection, and meet the definition 5 the place invariant branches set of the collection;

(6) Verification all elements in the invariable place subset collection correspond to the subnet is met Theorem 4 (6)-(8) condition. If so, you seek to create process; otherwise, end;

(7) Meet Theorem 4 (6)-(8) conditions of the invariable place subset collection corresponds to the subnet to P/T nets features division.

IV. PARALLEL ANALYSIS AND PARALLEL ALGORITHMS TO THE PROCESS OF PETRI NETS

A. *Multi thread parallel model of Petri net based on multi core PC*

Petri nets parallel is the third step in the process of the above method, Petri nets is functionally classification into separate processes; each process can be mapped to processors in parallel platforms and parallel programming of process[5]. As shown in Figure 1, that it is a multi thread model of Petri net based on multi core PC.

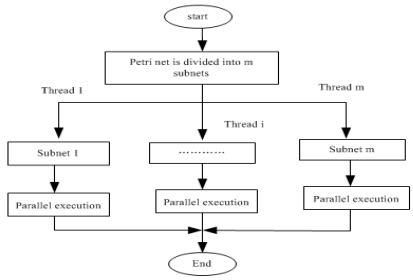


Fig.1 Multi thread model of Petri net based on multi core PC

B. *Analysis to internal transformer behave of process*

From figure 3, Petri nets after functional division, as an in-process place that contains the implementation of state or the state of resources, transition is a process to perform an action or operation. The following is analysis of the features of Petri nets process.

The first situation: If in the process of changing t, meet $|\cdot t|=|t\cdot|=1$, that transformer when only one input and one output, then this transformer is the process of local behavior or local action;

The second situation: If any two transformers in the process of t_1 and t_2, there is an identified M, M [t_1>M_1→M_1

[t_2 > and M [t_2>M_2→M_2 [t_1>, is in the process of transformer t_1 and t_2 are concurrent, The two transformers do not influence each other;

The third situation: If any two transformers in the process of t_1 and t_2, there is an identified M, M[t_1>M_1→¬ M_1[t_2 > and M[t_2>M_2→¬ M_2[t_1>, the process transformers in t_1 and t_2 conflict on M. Two transformers in conflict, if one occurs, another will lose rights, and vice versa. Address transitions conflicts, through the imposition of external control, joined the place p_a and p_b, $p_a t_1 p_b t_2$ form a control loop, t_1 and t_2 elimination of conflicts [7-9].

Realization of the function of the process must take into account the evolution of different situations. (1) If transformers belonging to first kind of situation, transformers of behavior is local behavior, in programming implementation, these behavior occurred in with a process internal, by process is located of processor serial implementation; (2) If transformer belonging to second kind of situation, these transformers can concurrent implementation, can used dispersed type work pool policy to implementation; (3) If transformers belonging to third kind of situation, process cannot normal of run. Join the two places so that they form a control circuit, t_1 and t_2 conflicts will disappear, forming two concurrent transformers, in which case the fall into second place.

C. Parallel analysis in the process

If the transformer is shared between the two processes, that is, $\Gamma_p \cap \Gamma_d = \{t_i\} \neq \varnothing$, two processes are transformers more than one input and output place, this two processes synchronization transformers. This transformer to the role of a server, through this server can extransition information between the two processes. When two processes are actually information extransition with a server;

Process concurrent with the establishment of Petri nets is directly related to practical problems. Some system processes are executing concurrently, some system processes are executed sequentially. This requires specific analysis of specific problems. In message-passing systems, if the process is executed concurrently, the parallel implementations, using asynchronous methods to achieve; if the processes are executed in order lines can be used to achieve.

D. Parallel algorithm

According to the analysis in the preceding section, with messaging system as the platform, Petri nets parallel algorithm is as follows:

Step1 Based on Petri nets function decomposition algorithm, to divide the nets into several processes;

Step2 According to the conditions $\Gamma_p \cap \Gamma_d = \{t_i\} \neq \varnothing$, identify shared transformer between processes, every shared transformer as a process server;

Step3 Map number and server processes (distribution) in different processor of a messaging system platform. Petri nets functions by the processor to complete the process, the server processes information transmitted between processes;

Step4 Petri functions of the process of implementation. Every process is a sequential state machine can be viewed as a programming task model, is reached when no successor state place, the loop end. Therefore, parallel implementation of message passing system achieves Petri nets system function;

Step5 Realize to functions of internal process transformer and conflict. According to 4.1, if more than two transformers are executed concurrently, decentralized work pool is used to achieve, these transformers as while running the local action, among the processors retain a transformer things, other transformers available to other processors to perform. In this way, and several child processes of this process, handled by the appropriate processor. If several transformers in conflict, then to add place, constitute a control loop after these transformers from a conflict to a concurrent execution, the process comes down to transformer concurrent processing;

Step6 Realization of the process of internal local action. Local action is the implementation of the transformer of procedure, simply order programming to internal process or handle the processor.

Step7 functional implementation of shared between two processes transformer server. Transformers of the function includes information received from two connected processes, information processing, and sending a message to connected two processes of local action and so on.

V. APPLICATION EXAMPLES

A. Converting the AMT Deposit and Withdrawal System into P/T Net Model

An simulation AMT bank deposit and withdrawal system (as shown in Figure 2) is used to examine the functional partitioning algorithm of Petri net system, and some specific parameters of this system are as follows:

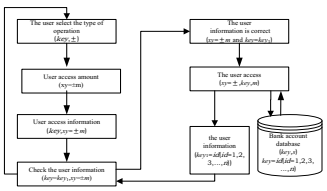


Fig.2 AMT access data stream

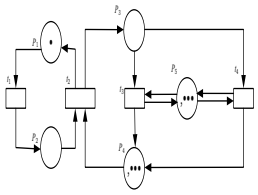


Fig.3 AMT in Petri nes

- key represents the user account, s refers to the deposit amount of users, s value may be different for different users;
- Database storage user information is (key,s), representing the current deposit amount of the user key in the bank;
- Operation-type character "+" indicates the deposit, "-" indicates the withdrawals, "±" indicates "+" or "-";
- Operation is represented by xy, xy=+ or xy=-

respectively represents the deposit or withdrawal operation;

■ （xy,key,m） represents the operation of m Yuan deposit or m Yuan withdrawal for the user with the account key;

■ xy=± m and key=key1 represents that the user account is the same as the bank account, then the deposit or withdrawal operation can be conducted for the user with the account key1;

■ Assuming that the number of users is *n*, *key*=*id*=1,2,3,…,*n*.

In order to solve the above practical problems, and according to the deposit and withdrawal data flow in Figure 2 and the above parameters, the actual system is converted into P/T net model, as shown in Figure 3. Wherein, p_1 indicates the user operation type, • represents that this place has one token; p_2 represents the user deposit and withdrawal information, p_3 indicates that the user information is correct, p_4 indicates that this place includes *n* user accounts and password information, stipulating that ", •••" is used to represent all *n* tokens of this place, instead of 3 tokens; similarly, p_5 represents the bank account database, ", •••" indicates all n tokens of this place; t_1 indicates the user deposit and withdrawal operation, t_2 represents receiving the user account and verifying the user information, t_3 represents the automatic deposit operation of the machine, and t_4 represents the automatic withdrawal operation of the machine.

B. Experimental Results and Analysis of AMT Deposit and Withdrawal System

(1) The experimental results are shown below in Table 1:

Table 1 Experiment Checking Results of AMT Deposit and Withdrawal System

Input data	The output results	Conclusion
D+ [4][5] = {{0, 1, 0, 0, 0}, {1, 0, 1, 0, 0}, {0, 0, 0, 1, 1}, {0, 0, 0, 1, 1}}; D−[4][5] = {{1, 0, 0, 0, 0}, {0, 1, 0, 1, 0}, {0, 0, 1, 0, 1}, {0, 0, 1, 0, 1}} M_0={1, 0, 0, *n*, *n*}	X_1=[1, 1, 0, 0, 0]^T, X_2=[0, 0, 1, 1, 0]^T, X_3=[0, 0, 0, 0, 1]^T Γ_{P1}={X_1, X_2}, Γ_{P2}={X_1, X_3} subnets=2	Can be divided into 2 subnets

(2) Analysis of experimental results: the subset Γ_{p1}={X_1、X_3} and Γ_{p2}={X_1、X_2} all satisfy the conditions of Theorem 2. Therefore, AMT deposit Petri net system has two kinds of function partitioning, the first is to divide into two processes p_a and p_b according to the corresponding subnet of X_1 and X_2. wherein, the function of process p_a is to operate the generating, namely, if the state of place p_1 is （"+",*key*,*m*）, then the transition t_1 is generated *xy*="+", and the state of place p_2 is (*xy*="+",*key*,*m*); if the state of place p_1 is ("-", *key*,*m*), then the transition t_1 is generated *xy*="-", and the state of place p_2 is (*xy*="-",id,n); the function of process p_b is to manage the operation, namely, when the state of place p3 is *xy*="+"and *key*=*key*$_1$, the *m* Yuan deposit operation is from the transition t_3, and when the state of place p_3 is *xy*="-"and *key*=*key*$_1$, the *m* Yuan withdrawal operation is from the transition t_4, and the second is to divide into two processes p_c and p_d according to the corresponding subnet of X_1 and X_3 (with similar functions).

VI. CONCLUSIONS

This article through to high-level Petri nets automatically converted to P/T nets, using the invariants place in Petri nets system process function division, and expanse the division condition, get the invariants place function partition and parallel calculation method of Petri nets system. Finally, the analysis and checking for a typical bank deposit and withdrawal Petri net system instance. Examples show, Petri nets system for automatic conversion, invariants place function partition and parallel calculation method is feasible and effective. The algorithm will be the future of our research work focus in areas related to specific application and simulation.

REFERENCES

[1] M. Paludetto. Sur la commande de procedes industriels:unemethodologie basee objets et reseaux de Petri. These de doctorat,Universite Paul Sabatier,Toulouse, France,1991 : 34-47.

[2] W.El Kaim and F.Kordon. An integrated framework for rapid system prototyping and automatic code distribution. In 5thIEEE International Workshop on Rapid System Prototyping,Grenoble,IEEE Comp.Soc.Press,1994:52-61.

[3] J.M.Colom and M.Silva. Convex geometry and semiflows in P/T nets. A comparative study of algorithms for computation of minimal P-semiflows. In Rozenberg,Advances in Petri Nets 1990,Volume 483 of Lecture Notes in Computer Science.Springer-Verlag,1991:79-112.

[4] WU Z H.Petri Nets Introduction[M]. Beijing: Mechanical industry publishing house.2006, pp.144-155.

[5] Girault C and Valk R.Petri Nets for Systems Engineering: A Guide to Modeling, Verification, and Applications [M].Springer-Verlag Berlin Heidelberg. 2003,pp.159-235.

[6] Wenjing LI, Shuang LI, Zhong-ming Lin. Research on Petri nets parallelization the functional divided conditions. 12th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Publicshed by IEEE conference publishing services ,2013,pp.50-54.

[7] J. M. Couvreur, D. Poitrenaud and P. Weil. Supporting Processes of General Petri Nets[C]. In proceeding of: Applications and Theory of Petri Nets - 32nd International Conference, PETRI NETS 2011, Newcastle, UK, June 20-24, 2011: 129-148.

[8] V. Khomenko and A. Mokhov. An Algorithm for Direct Construction of Complete Merged Processes[C]. 32nd International Conference, PETRI NETS 2011, Newcastle, UK, June 20-24, 2011:89-108.

[9] R. Kocí and V. Janoušek. Towards Design Method Based on Formalisms of Petri Nets[C], DEVS, and UML. In ICSEA 2011, The Sixth International Conference on Software Engineering Advances, 2011:299–304.

GRIB Parallel Design of Civil Aviation Meteorological Data Processing System

Zhengwei Guo, Yongwei Gao, Yafei Jiang, Guang Xue
School of Computer and Information Engineering
Institute of Image Processing and Pattern Recognition
Henan University
Kaifeng, Henan, 475001, China
Email: gzw@henu.edu.cn, 2386113282@qq.com, henugao@163.com, xueguang@henu.edu.cn

***Abstract*—With the increasing volume of business the civil aviation data processing system, the data processing efficiency requirements become higher and higher. In order to improve the speed of data processing, we design a parallel scheme of GRIB according to the GRIB data processing flow. We achieved the parallel processing of GRIB data reading and decoding process, and carried out the related test. It is significantly improve the speed of data loading.**

Keywords-Civil Aviation Meteorological Data Processing System; GRIB; Multithreading; Parallel design.

I. INTRODUCTION

The meteorological data code mainly includes character code and Form Code in the operational meteorological services. The character code is simple and intuitive. Every coded must have an appropriate decoding process. It is difficult to meet the growing meteorological data format. The characteristic of table-driven code is self-describing capability, scalability, data compression strong ability, code simplification of procedures. As a result, the WMO (World Meteorological Organization) recommends gradual change from the character encoding format to a table-driven coded [1]. GRIB（GRIded Binary）data uses the table-driven code of Binary Grid data processing code. GRIB code divided into two categories GRIB Edition 1 and GRIB Edition 2 in the international guidelines, and main storage the values of meteorological [2]. In the civil aviation meteorological data processing system, the 12 kinds of meteorological data processing are processed in parallel, and the 12 kinds of meteorological data have been achieved in parallel. It takes a lot of time to process large amounts of data in the GRIB file processing system. Based on the GRIB data processing features, adopt the parallel design of the GRIB grid point data in the existing system version to improve GRIB data processing speed.

II. PROCESS FLOW OF GRIB

The process flow of GRIB as shown in Figure 1.

GRIB decoding process is divided into three sections. First, start up the listener and process of GIRB. It automatic monitors the GRIB data file which is located in the /data/grib directory. Then when the communication data stored in the directory of /data/grib, the GRIB process to read a group of GRIB data file list, starting on the list of documents one by one to decode and storage. Finally, save the copy files after a data file processing is completed, at this point the data files of /data/grib directory are deleted, and this one GRIB file is processed.

GRIB1 decoding process is divided into 6 sections. The section 0 is the starting point for the decoding. It mainly obtained the length of the entire data and the version number of GRIB. In the section 1, the length of section, version number of the parameter table, data processing center identification process ID number and identifier are obtained. Section 2 is an option, according to the presence or absence of coded to obtain the length of section and geometry described. Section 3 is also an option, according to the presence or absence of coded to obtain the length of section and bit-map. In the section 4, the length of section and the parameters E, X, and R of

$$Y_K \cdot 10^D = R + X_K \cdot 2^E \ (K=1, 2, 3 \ldots) \quad (1)$$

are obtained[3]. In the section 5, it gets the ASCII code represented by "7777" marks the end.

GRIB2 decoding process is divided into 9 sections, section 0 is the same as GRIB1 obtained the entire length and GRIB version. In the section 1, the identifier of prediction, the length and coding analysis are obtained. Compared with the GRIB1, the characteristics of processing data are increased. Although the section 2 is an option, it is also important. This section usually required to send in messages. It obtains the length, section number and other information by decode, and mainly local information is added by the filer center. The information obtained in the section 3 is different from GRIB1. This section is not option, and obtains the geometry of grid data, the length and section number [1]. In the section 4, it obtains the length, the character of the data and section number by decode, which mainly describes the character of data. In the section 5, it obtains the length, section number and description representation of data values. Three templates are used in this section of data values. Different templates are selected according to the different values. In the section 6, it contains the data bit-map and the length of this section. The storage format is a bit of a grid can only be stored in a sequence. It should be noted that values are only 0 and 1, and 0 represents the position of grid data is omitted, 1 represents the position of grid data has not been omitted, bit value and grid data bits correspond to each other. In the section 7, it main obtain the length of section, section number and grid information. In the section 8, it gets ASCII code represented by "7777" marks the end.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.16

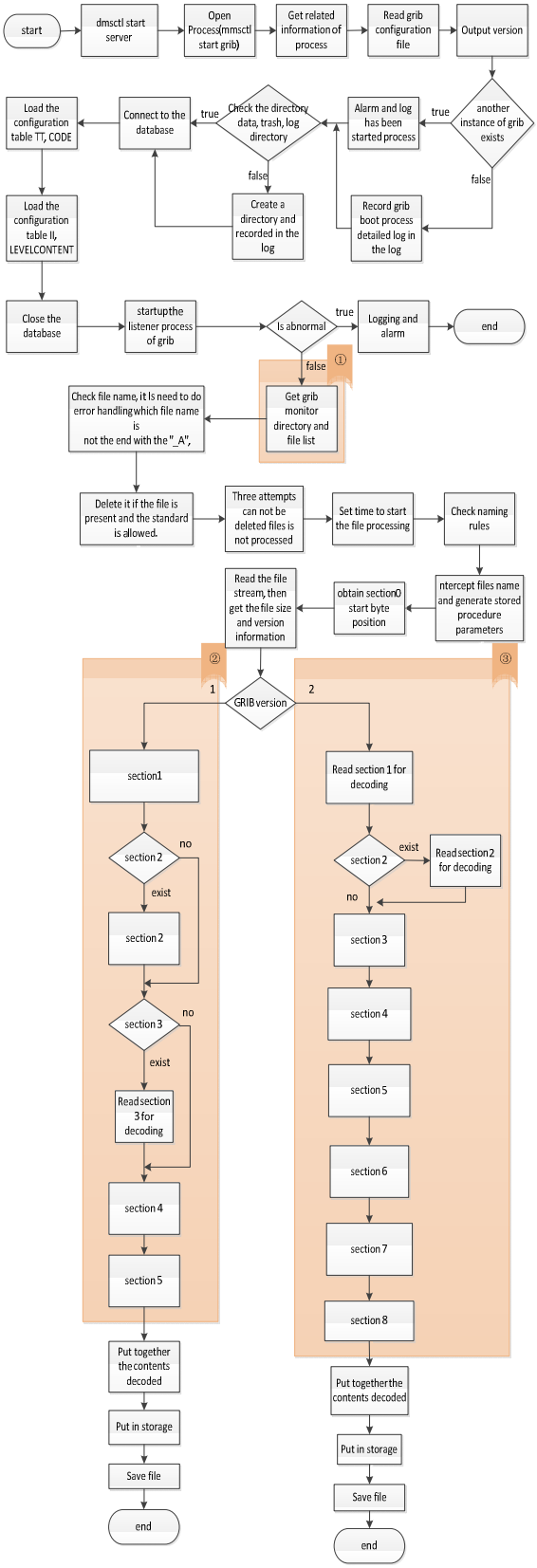


Figure 1. GRIB process flowchart.

III. PARALLEL SCHEME AND PARALLEL DESIGN

A. *Parallel Scheme*

There are two places can be improved in the process of GRIB processing:

The first, a problem of process the files after get the GRIB data file list. Now the project file list processing is handled one by one. So there is room for improvement on the efficiency of the process.

The second, in the GRIB decode process, now the project of GRIB is decoded in a serial way. The decode process also has a space to improve the efficiency here.

After analyze the GRIB process flow, it can implement the parallel design to improve the speed of data processing in the area of ①②③ as shown in Figure 1. In the area of ①, files are processed one by one instead of process the files in batches after read the GRIB data directory. In the area of ② and ③, the decode is carried out in serial mode after break the work of decode instead of multi-threaded parallel execution[4].

B. *Parallel Processing in Batches*

The multiple files are processed in the area of ①. Due to the GRIB file process is too complex and each message only extract a file to process from the directory of data/grib. There is a part of GRIB files accumulated in data/grib directory. As a result of the speed of process is relatively slow. A parallel scheme is implemented here. The processing logic is refreshed every 500 milliseconds. Extract the list of files in the directory when the large file in the data/grib directory. When the file list is finished, the number of threads is started up according to the hardware CPU kernel. The number of files obtained in the file list is assigned to the corresponding thread [5]. As shown in Figure 2, if the hardware of the CPU kernel number is n, the program corresponding to startup n threads. If there are m files in the list of files, then the file is allocated to m/n and takes an integer. The remainder is assigned to the last thread.

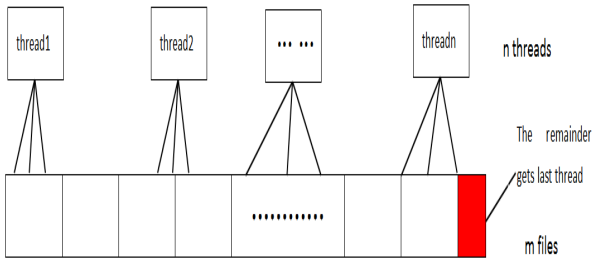


Figure 2. Assignment of thread tasks

C. *Grib Files Internal Segment Parallel Processing*

In the area of ②③ as shown in Figure 1, the standard of GRIB grid data is divided into two categories: GRIB1 and GRIB2. According to the decoding of GRIB1, it is divided into 6 sections, and the data storage program is also divided into 6 sections.

It is GRIB1 decoding process in the area of ②. First read the beginning of the message, and then complete decoding section 0. After the section 0 decoding is completed, read the section 1 and decoding. After the section 1 is processed, then read the flag of section 2 and section 3. If both the section 2 and section are exist, decoding the section 2 in front of the section 3. If the section 2 does not exist, then judge whether the section 3 exists. If the section 3 does exist, it processes the section 3. If the section 3 does not exist, it processes the section 4. Finally execute the section 5. In this point, if each segment is extracted, a parallel implementation of the decoding can be used as a parallel way [6].

Because of the section 2 and section 3 are required to judge whether exist, so section 1, section 2 and section 3 need to be assigned in a same thread, the other sections are assigned one thread, it is divided into four threads.

GRIB1 decoding process threads allocation:

Thread1: section 0.

Thread2: section 1, section 2, section 3.

Thread3: section 4.

Thread4: section 5.

It is GRIB2 decoding process in the area of ③. According to the decoding of GRIB2, it is divided into 9 sections, and the data storage program is also divided into 9 sections. Now running in the decoding is achieved in a serial way. Several related templates are used in the section 5 and section 7 of GRIB2. The section 5 is divided into 3 templates, which template is decided by the value. The section 7 is divided into 4 templates, which template is decided by the value. It takes a lot of time to process the template. GRIB2 decoding process can be designed as multi thread parallel method.

The section 2 is required to judge whether exist. When design a multi thread, it is should to be assigned a thread to section 1 and section 2, the other sections are executed by one thread alone, and it is divided into 4 threads. After the thread executes the decoding process to determine the number of times the thread is executed, if the thread1, thread 3 and thread 4 threaded decoding number is two times, the thread 2 decoding number as same as the expected time. Then the decoding meets requirements. A decoding process is divided into four parallel to achieve the decoding thread.

GRIB2 decoding process threads allocation:

Thread1: section 0, section 5.

Thread2: section 1, section 2, section 6.

Thread3: section 3, section 7.

Thread4: section 4, section 8.

IV. GRIB MULTITHREADING PARALLEL TEST

After GRIB paralleled, then carry out the data test.

Test environment: Database Server: AIX, Database: DB2. Application server: 4-core CPU. Open 4 threads. The Test data of GRIB is meteorological data. Some day's all real-time data of GRIB.

The testing scheme is divided into three: Parallel testing of only multi files in batches; parallel testing of the file's internal segmentation; testing for the parallel test of file's internal segmentation and multi file. The following table is the result of the test.

The parallel testing result of only multi files in batches, Table I:

TABLE I. THE PARALLEL TESTING RESULT OF ONLY MULTI FILES IN BATCHES

Parameter	Before Parallel Processing	After Parallel Processing
Number of files	800	800
CPU Usage Rate (%)	about 60%	about 300%
Spend time	13min	6min
Number of files	5000	5000
CPU Usage Rate (%)	about 60%	about 300%
Spend time	1h 35min	42min
Number of files	25507	25507
CPU Usage Rate (%)	about 60%	about 300%
Spend time	7h 5min	2h 50min

Parallel testing results of the file's internal segmentation are listed in Table II. Parallel testing results of the file's internal segmentation and multi file in batches are listed in Table III.

TABLE II. PARALLEL TEST RESULTS OF THE FILE'S INTERNAL SEGMENTATION

Parameter	Before Parallel Processing	After Parallel Processing
Number of files	25507	25507
CPU Usage Rate (%)	about 60%	about 380%
Spend time	7h 5min	5h 10min

TABLE III. PARALLEL TEST RESULTS OF THE FILE'S INTERNAL PARTITION AND MULTI FILE IN BATCHES

Parameter	Before parallel processing	After parallel processing
Number of files	25507	25507
CPU Usage Rate (%)	about 60%	about 390%
Spend time	7h 5min	4h

According to the result of test, the parallel testing of only multi files in batches is the best. Blocking decoding is not the main time of process and parallel decoding also can improve efficiency, but it is not better than the parallel test of multiple files in batches. If make the parallel and decoding together，test 25507 data, the usage of CPU is too high to reduce the processing speed. In the decoding process, it is required to wait for other threads when splice information, so it wastes the some execution time. In fact, during the realization of multi file sub batch processing, it has reached more than one file to decode the work at the same time. So it can be regarded as the decoding of parallel. Finally, the parallel of multiple files in batches is more appropriate at here.

V. CONCLUSION

After the analysis of the process flow of the GRIB file carefully and several tests. The concurrent design is feasible. In the civil aviation meteorological data processing system, through parallel design the processing efficiency of the GRIB file has been improved obviously. After the hardware reaches a certain requirement, the parallel of the GRIB grid point data has a great improvement in the time and resource utilization, so that the system can provide services more efficiently.

REFERENCES

[1] National Meteorological Information Center. Table-driven coding guide [M] Beijing : Meteorological Press, 2005:9-34.

[2] Word Meteorological Organization. Introduction to GRIB Edition 1 and GRIB Edition 2.

[3] Sun Xiubin, Ying Xianxun. GRIB code compression principle and algorithm [J]. Meteorological Science and Technology , 1992,18(9):46-48.

[4] Guo Guangjun, Hu Yuping, Dai Jingguo. Research and application of parallel computing technology based on Java multi thread [J]. Journal of Central China Normal University.2005.6:169-173.

[5] Peter S.Pacheco. An Introduction to Parallel Programming [M] Beijing : Machinery Industry Press, 2011.9.

[6] Liu Tao, Fan Bin, Wu Chengyong, Zhang Zhaoqing. Data stream Java programming model of parallel design, implementation and optimization of runtime [J].2008.9:2181-2190.

A parallel algorithm of Green Function with Free Water Surface

Chao Sun[1], She-sheng Zhang[2]
[1] Zhixing College of Hubei University, Wuhan, P. R. China
[2] Wuhan University of Technology, Wuhan, P. R. China.
10368260@qq.com

***Abstract*—According to the convenient using principle of parallel computation, the control equation and boundary condition of point source with free water surface are given, the basic analyzing solution is obtained, the Green function representation is discussed., the discrete calculation expression and calculation procedure are proposed with parallel computer, two-dimensional graphics of the Green function's real and imaginary part are plotted.**

Keywords: Green function, free surface, parallel computer

I. INTRODUCTION

In the case of calculating add mass spend a lot CPU time, the parallel algorithm is consider in the field of ship research. Add mass is calculated usually by using boundary element method. The kernel function expression form of the boundary integral is complicated, if the water free surface is considered. The kernel function is called Green function in the case of that the singular point is point source. The Green function with free surface condition has two singular integral points, which leads to huge calculation and larger error. With the rapid improvement of the computer software and hardware, lots of scholars take researches on the Green function. Huan[1] proposes that concisely and precisely obtains the time-domain Green function and its spatial derivative is the key of the ship hydrodynamics problems. Author of Han[2] uses multi-dimensional polynomial approximation to replace the direct numerical calculation with integral form, and this replacement can be adopted in the boundary element method calculation. Dan[3] takes some researches on the two-dimensional time-domain Green function and its partial derivatives calculation in ship hydrodynamic issues, proposes a new expression form with these two functions, and a new creating table interpolation method to obtain function results. Xie[4] takes integration on the limited depth complex Green function, and the result agrees well with the result from the Norway DNV Classification Society SEAM software. Liu[5] separate the part got by direct calculation from the integral function by reduction of a fraction. This procedure reduces the order of the left part in the integral function and reduces the calculation by using the unlimited depth Green function theory. Xie[6] combine the three-dimensional potential flow theory with limited depth complex Green function, and calculates the water elastic response of the Floating Production Storage & Offloading(FPSO). Shen[7] proposes the ordinary differential equations about depth Green function and its derivative, and a rapid Green function calculation method combining solving ordinary differential and interpolation between nodes. Liu[8] proposes a convolution calculation recurrence formula by using the Fourier transform relation between time-domain Green function and frequency-domain Green function[9]. This method largely reduces the calculation difficulty[10].

Because of the rapid development in parallel computing[11], it is necessary to consider the calculation method of the Green function on the parallel computing platform[12]. This paper takes the research on the parallel calculation of the Green function with free surface[13].

II. TWO DIMENSION GREEN FUNCTION

Suppose velocity potential φ satisfies Laplace equation

$$\Delta\phi = \delta(P-Q) \tag{1.1}$$

Here P is field point Z=x+iy，Q is source point ζ = ξ +i η .The right of equation is delta function. It can be rewritten as:

$$\frac{\partial^2\varphi}{\partial y^2}+\frac{\partial^2\varphi}{\partial x^2}=\delta(x-\xi)\delta(y-\eta)$$

Here dealt function is defined as

$$\delta(x-\xi)=\begin{cases}\infty & x=\xi\\ 0 & x\neq\xi\end{cases}$$

And

$$\delta(y-\eta)=\begin{cases}\infty & y=\eta\\ 0 & y\neq\eta\end{cases}$$

The boundary condition is：

$$\frac{\partial\varphi}{\partial y}-k\varphi=0, y=0 \qquad -\mathrm{Im}(\frac{d\Phi}{dz})-k\,\mathrm{Re}\,\Phi=0$$
$$\frac{\partial\varphi}{\partial x}\pm ik\varphi=0 \qquad x=\infty \qquad \mathrm{Re}(\frac{d\Phi}{dZ})\pm ik\,\mathrm{Re}\,\Phi=0 \tag{1.2}$$

Here Φ is complex potential. On the free surface y=0, velocity potential satisfies linear condition. The analysis solution is called green function as:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.17

$$G(z,\zeta)=\frac{1}{2\pi}\ln\frac{Z-\zeta}{Z-\bar{\zeta}}-\frac{1}{\pi}I+i\exp(-ik(Z-\bar{\zeta})$$

$$I=PV\int_0^{\infty}\frac{\exp(-iu(Z-\bar{\zeta})}{u-k}du$$

(1.3)

Here I is principle value integration and u=k is zero point of denominator. The integral domain is infinite. From the expression of integration, the value I is determined by the parameters of k,(x- ξ),(y+ η). It can be rewritten as when k=1:

$$G(z,\zeta)=\frac{1}{2\pi}\ln\frac{Z-\zeta}{Z-\bar{\zeta}}-\frac{1}{\pi}I+i\exp(-i(Z-\bar{\zeta})$$

$$I=PV\int_0^{\infty}\frac{\exp(-iu(Z-\bar{\zeta})}{u-1}du$$

Let u/k-->u, Z-->kZ, we have

$$G(z,\zeta)=\frac{1}{2\pi}\ln\frac{kZ-k\zeta}{kZ-k\bar{\zeta}}-\frac{1}{\pi}I+i\exp(-i(kZ-k\bar{\zeta})$$

$$I=PV\int_0^{\infty}\frac{\exp(-iu(kZ-k\bar{\zeta})}{u-1}du$$

If source point $\zeta=\xi$ is on the free water surface, the Green function may written as:

$$G(z,\zeta)=-\frac{1}{\pi}I+i\exp(-ik(x-\xi+iy)$$

$$I=PV\int_0^{\infty}\frac{\exp(-iu(x-\xi+iy)}{u-k}du$$

Or

$$G(z,\zeta)=-\frac{1}{\pi}I+i\exp(-ik(x-\xi+iy)$$

$$I=PV\int_0^{\infty}\frac{\exp(-iu(kx-k\xi+iky)}{u-1}du$$

III. The Representation of Theoretic Formula

On the parallel computing platform, it is required simple and clear to express numeric formula. Let unit transfer as

$$X=-i(Z-\bar{\zeta})/R\qquad R=|Z-\bar{\zeta}|$$

(2.1)

The principle value integration may be written as

$$I(a)=\int_0^{\infty}\frac{\exp(uRX)}{u-k}du=H(kR,\theta)$$

(2.2)

Where H(a,b) is two parameters function defined as:

$$H(a,\theta)=\int_0^{\infty}\frac{\exp(vX)}{v-a}dv$$

(2.3)

And v=uR,

$$X=-ie^{i\theta}=-e^{i\delta}$$

$$\delta=\theta-1.5\pi\qquad -0.5\pi\le\delta\le 0.5\pi$$

(2.4)

Let s=v-a, we have

$$H=e^{aX}[C_a-\sum_{n=1}\frac{(-aX)^n}{n!n}-\log(a)]$$

(2.5)

We have representation for small abstract value of X approximately:

$$H=e^{aX}[C_a+aX-\log(a)]$$

Second order is

$$H=e^{aX}[C_a+aX-\frac{(aX)^2}{4}-\log(a)]$$

Third order is

$$H=e^{aX}[C_a+aX-\frac{(aX)^2}{4}+\frac{(aX)^3}{18}-\log(a)]$$

From above results, we find the convergence speed is fast for series.

We have

$$H=e^{a(x+iy)}[C_a-\sum_{n=1}\frac{(-a)^n e^{in\theta}}{n!n}-\log(a)]$$

$$=e^{ax}[\cos(ay)+i\sin(ay)]$$

$$*[C_a-\sum_{n=1}\frac{(-a)^n\cos(n\theta)}{n!n}-i\sum_{n=1}\frac{(-a)^n\sin(n\theta)}{n!n}-\log(a)]$$

The real and image part are:

$$\mathrm{Re}\,al(H)=e^{ax}\{\cos(ay)[C_a-\sum_{n=1}\frac{(-a)^n\cos(n\theta)}{n!n}-\log(a)]$$

$$+\sin(ay)\sum_{n=1}\frac{(-a)^n\sin(n\theta)}{n!n}\}$$

$$image(H)=e^{ax}\{\sin(ay)[C_a-\sum_{n=1}\frac{(-a)^n\cos(n\theta)}{n!n}-\log(a)]$$

$$-\cos(ay)\sum_{n=1}\frac{(-a)^n\sin(n\theta)}{n!n}\}$$

First order of H is

$$\mathrm{Re}\,al(H)=e^{ax}\{\cos(ay)[C_a+a\cos(\theta)-\log(a)]$$

$$+a\sin(ay)\sin(\theta)\}$$

$$image(H)=e^{ax}\{\sin(ay)[C_a-a\cos(\theta)-\log(a)]$$

$$-a\cos(ay)\sin(\theta)\}$$

Second order of H is

$$\mathrm{Re}\,al(H)=e^{ax}\{\cos(ay)[C_a+a\cos(\theta)-\frac{(a)^2\cos(2\theta)}{2}$$

$$-\log(a)]+\sin(ay)[-a\sin(\theta)+\frac{(a)^2\sin(2\theta)}{2}]\}$$

$$image(H)=e^{ax}\{\sin(ay)[C_aC_a+a\cos(\theta)-\frac{(a)^2\cos(2\theta)}{2}$$

$$-\log(a)]-\cos(ay)[-a\sin(\theta)+\frac{(a)^2\sin(2\theta)}{2}]\}$$

The Ca is

$$C_a=\frac{1}{X}\int_0^1 e^{X/y}dy-\sum_{n=2}\frac{X^{n-1}}{n!(1-n)}-1-\frac{1}{X}$$

(2.6)

We have linear order

$$C_a=\frac{1}{X}\int_0^1(1-\frac{X}{y})dy-X-1-\frac{1}{X}$$

It is easy to find that Ca is the function of parameter X(or δ), and independent with parameter a.

IV. The Parallel Computation of Numeric Steps

On the parallel computing platform, it is required that the numeric steps is related with numeric formula, and

easy used to write code. The numeric steps of Green function are:

(1) Determine the domains D of parameters a and δ;

(2) Choose N processors to parallel calculate the value of function H(a, δ);

(3) Divide numeric domain D to N sub-domains.

(4) Calculate H value on sub-domain;

(5) exchange data with master computer;

(6) Output numeric result.

When a=1, we have

$$H = e^{(x+iy)}[C_1 - \sum_{n=1} \frac{(-1)^n e^{in\theta}}{n!n}] \quad (3.1)$$

$$= e^{x}[\cos(y) + i\sin(y)]$$

$$*[C_1 - \sum_{n=1} \frac{(-1)^n \cos(n\theta)}{n!n} - i\sum_{n=1} \frac{(-1)^n \sin(n\theta)}{n!n}]$$

The real and image part are:

$$\mathrm{Re}\,al(H) = e^{x}\{\cos(y)[C_1 - \sum_{n=1} \frac{(-1)^n \cos(n\theta)}{n!n}]$$

$$+\sin(y)\sum_{n=1} \frac{(-1)^n \sin(n\theta)}{n!n}\}$$

$$image(H) = e^{x}\{\sin(y)[C_a - \sum_{n=1} \frac{(-1)^n \cos(n\theta)}{n!n}]$$

$$-\cos(y)\sum_{n=1} \frac{(-1)^n \sin(n\theta)}{n!n}\}$$

When a= π , we have

$$H = e^{\pi(x+iy)}[C_\pi - \sum_{n=1} \frac{(-\pi)^n e^{in\theta}}{n!n} - \log(\pi)]$$

$$= e^{\pi x}[\cos(\pi y) + i\sin(\pi y)]$$

$$*[C_\pi - \sum_{n=1} \frac{(-\pi)^n \cos(n\theta)}{n!n} - i\sum_{n=1} \frac{(-\pi)^n \sin(n\theta)}{n!n} - \log(\pi)]$$

The real and image part are:

$$\mathrm{Re}\,al(H) = e^{\pi x}\{\cos(\pi y)[C_\pi - \sum_{n=1} \frac{(-\pi)^n \cos(n\theta)}{n!n} - \log(\pi)]$$

$$+\sin(\pi y)\sum_{n=1} \frac{(-\pi)^n \sin(n\theta)}{n!n}\}$$

$$image(H) = e^{\pi x}\{\sin(\pi y)[C_\pi - \sum_{n=1} \frac{(-\pi)^n \cos(n\theta)}{n!n} - \log(\pi)]$$

$$-\cos(\pi y)\sum_{n=1} \frac{(-\pi)^n \sin(n\theta)}{n!n}\}$$

V. NUMERIC RESULTS

The numerical results are obtained by using parallel calculating. The number processor N=32. The parameter a is chosen from 0.1 to 1000. The parameter δ is chosen from 0.1 to 1.5; after obtain numeric results, the value of Green function is used to draw figure. Fig.1 shows the real of Green function varied with angle δ, in the figure, the real line isδ=-1.428， point real line is δ =-0.8568， dashed line is δ =-0.5712. Fig.2 shows the real part of H function varied with image part of H function. The parameter δ =-1.4708,-1.3708,-1.2708.

When δ =0,

We have θ =1.5 π ,and

$$H = e^{a(x+iy)}[C_a - \sum_{n=1} \frac{(-a)^n e^{in1.5\pi}}{n!n} - \log(a)]$$

$$= e^{ax}[\cos(ay) + i\sin(ay)]$$

$$*[C_a - \sum_{n=1} \frac{(-a)^n \cos(n1.5\pi)}{n!n} - i\sum_{n=1} \frac{(-a)^n \sin(n1.5\pi)}{n!n} - \log(a)]$$

$$= e^{ax}[\cos(ay) + i\sin(ay)]$$

$$*[C_a - i\sum_{n=1} \frac{(-a)^n \sin(n1.5\pi)}{n!n} - \log(a)]$$

VI. CONCLUSION

There is hug computation of Green function with free water surface for ship hydrodynamics. The paper constructed the control equation and boundary condition of point source with free water surface by using parallel computation method, the form of analyzing solution is obtained, the Green function representation is discussed., the discrete calculation expression and calculation procedure are proposed with parallel computer, two-dimensional graphics of the Green function's real and imaginary part are plotted.

ACKNOWLEDGMENT

The paper is financially supported by by China national natural science foundation (No.51139005),

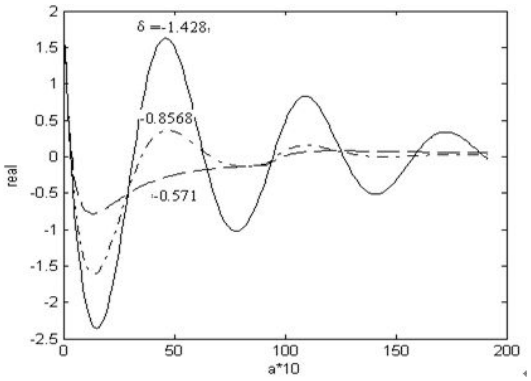


Fig.1 the real of Green function varied with angle δ

REFERENCES

[1] Huan Debo, time domain green function and its derive numeric calculation[J], China Ship building, 1992(4),1

[2] Han Ling, Teng Bin, Guo Ying, Approximation of time domain Green function[J], J of hydrodynamics(A),2004(5), 929-637.

[3] Dan Wenyang, Dai Yishang, Numerical evolution of two dimensional time domain green function[J], J of hydrodynamics,1996(6), 331-335

[4] Xie Yonghe, et all, Numerical calculation of finite water depth composite green function[J],J of Ship Mech.,2005(1), 23-28

[5] Liu Ri-rrdng,Ren Hui-long,Li Hui, A nimproved Gauss-Laguerre method for finite water depth Green function and its derivatives[J], J Ship Mech.,2008(2),p188-197.

[6] Xie Yonghel， et al, The Effects of Water Depth on Hydroelastic Response of a Very Lager FPSO[J], J Shanghai Jiaotong Univ., 2006(6), p993-997

[7] Shen Liang， et al, A practical numerical method for deepwater time domain Green function[J], J of hydrodynamics(A), 2007(3),380-386

[8] Liu Changfeng, et al, New convolution algorithm of time−domain Green function[J], J of hydrodynamics(A),2010(4), 25-34.

[9] Xin Chen, XinCong Zhou, Shesheng Zhang，Dan Li，A numerical fluid-solid coupling model for the dynamics of ships in atrocious sea conditions［J］, Journal of Algorithms & Computational Technology,2015,Vol. 9 No. 2， 163-175.

[10] Zixiang Yu, Dan Li, Jin Shengping, Yufeng Gui , Zhang Shesheng, The Parallel Computation of Green Function Based On the Characteristic Length of Ship[C], Proceeding of DCABES 2014 , 24~27OCT 2014 , Xianning, China. , pp168-170.

[11 Xin Chen, Dan Li, Shesheng Zhang,Computing Green's Function for the Free Water Surface Near Ship with large parameter by using parallel computer[C], Proceeding of DCABES 2014 , 24~27OCT 2014, Xianning, China., pp188-190.

[12] Xin Chen, Hualing Zhao, Yufeng Gui, Shesheng Zhang, Parallel numerical model of water lubricated rubber bearing[C] ,Proceeding of DCABES 2014, 24~27OCT 2014, Xianning, China., pp179-182.

[13] Xin Chen, Quan Kuang, Yang Li, Songbo Wang, Yunling Ye, Shesheng Zhang, Weis-Fogh Mechanism Mathematic Model of Wave Power Generation Device[J], Open Journal of Fluid Dynamics,Vol.4 No.4 2014,373-378

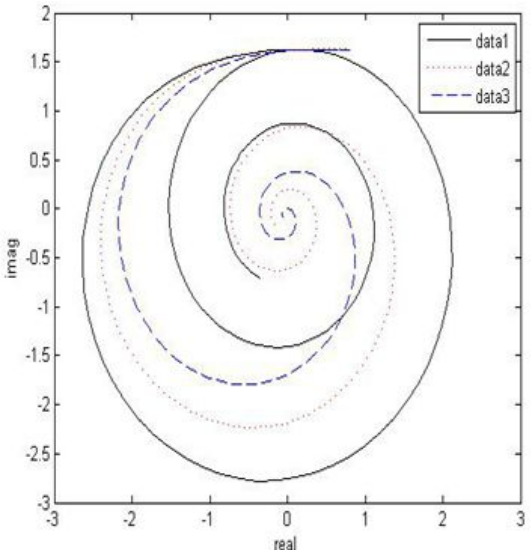


Fig.2 real part of H function varied with image part

Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm

Ding Yanrui, Zhang zhen, Wang Wenchao, Cai yujie
School of Digital Media
Jiangnan University
Wuxi, P. R. China
e-mail: yr_ding@jiangnan.edu.cn

***Abstract*—Modularization on the metabolic network can help to determine the relationship between community in network and network stability and evolutionary process. In this paper, we selected seven kinds of thermophiles and 7 kinds of mesophiles as the research objects, and constructed their metabolic networks using Pajek algorithm. Next, "component" definition and Girvan-Newman algorithm are used to identify the communities in the metabolic networks. The results showed that ratios of module number to node number are 15.71% and 16.90% respectively in thermophiles metabolic networks, while ratios of module number to node number are 17.61% and 19.79% respectively in mesophiles metabolic networks. The effects of these two methods of modularization show that modular degree in thermophiles is higher than in mesophiles. The minimum of Q function is 0.88, which means the performance of Girvan-Newman algorithm is better to identify communities. In addition, from the number of nodes in communities, we can deduce that the density in thermophilic bacteria metabolic network is larger than in mesophilic bacteria metabolic network.**

Keywords-mesophile; thermophile; metabilic network; modularization

I. INTRODUCTION

With further study on complex network properties, it's found that many real-world networks have a common character, i.e. community. The community in complex network is considered as substructure of the network. Connections between nodes in same community are intensive, while that is sparse between any different communities. In general, the community in the real-world network corresponds to some kind of natural phenomenon. For instance, realistic loading structures are represented by the community of the social network in terms of hobbies or backgrounds; the community of the citation network indicates same or similar subject of paper; Communities in the metabolic network are the representation of a life cycle or some modules with specific functions[1].

The metabolic network as well as other biochemical networks has complicated network properties, such as scale-free, small-worldness[2] and power-law distribution of nodes[3]. Moreover, the principle of the definition of modules in the metabolic network is that intra-module connections are intensive meanwhile inter-module connections are spares, has a certain structure of independence[1, 4-6].

To investigate the functional information contained in the metabolic network, it's necessary to identify the functional modules in it[7]. Researchers have proposed many effective methods to detect the communities in the metabolic networks. The spectral method that used to deal with image segmentation, has been applied to modularize complex networks recently[8]. It is based on strict mathematical theory, but needs some priori knowledge and can't ensure optimal network division. According to maximum flow-minimum cut of graph theory, Flake et al. proposed complex networks clustering algorithm—Maximum Flow Community (MFC)[9]. The Efficiency of MFC depends on the time of calculating minimum cutsets, it will takes a long time. Palla et al. suggested a clustering algorithm based on k-clique[5]. In practice, the value of parameter k significantly affect the module structure obtain, leads to obtain inconclusive result of modular structure. Girvan and Newman proposed GN algorithm inspired by definition of modularization and betweenness of edge, which provide an evaluation function without priori knowledge to judge modularity of networks[10]. Girvan and Newman first discovered cluster structure in networks, and many other clustering methods, such as hierarchical clustering in terms of similarity and network division based on the definition of component[11]. In addition, from the perspective of pattern recognition, Zhang et al. selected improved k-means algorithm(IKM) and improved fuzzy C-means clustering algorithm[12].

In this paper, we choose 14 typical biological samples from database of Ma Hongwu et al[13]. as research object. Using GN and modular methods based on 'component' which helps to find difference between thermophiles and mesophiles, we can identify thermotolerance factors that hide in networks of thermophiles.

II. DATA AND METHODS

A. dataset

We selected 7 thermophiles and 7 mesophiles as research objects based on growth temperature of microorganism in the dsmz database (http://www.dsmz.de/). We constructed metabolic networks of these 14 microorganisms with Pajek algorithm. The information of research objects is showed in table 1.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.18

TABLE I. Research Objects

	Microorganism name	Growth temperature(°C)
Thermophile(Bacteria)	A.aeolicus (aae)	95
	T.maritima (tma)	80
	T.elongatus (tel)	50
	T.engcongensis (tte)	75
Mesophile(Bacteria)	E.coli (ecc)	37
	S.coelicolor (sco)	28
	A.tumefaciens (atc)	26
	B.halodurans (bha)	30
Thermophile(Archaebacteria)	P.abyssi (pab)	97-100
	P.furiosus (pfu)	97-100
	M.kandleri (mka)	98
Mesophile(Archaebacteria)	H.sp.NRC-1(hal)	37
	M.acetivorans (mac)	37-40
	M.mazei (mma)	37

The living temperatures of microorganisms in table 1 from 26 °C to 100°C cover the entire thermophiles and mesophiles growth temperature range, So the sample we chosen is fairly typical. It could be discovered the network topology factors related to growth temperature of thermophiles and mesophiles by delving into metabolic networks communities.

B. Method of modularization

Modularization is a division of a network module. It is calculated as follow:

$$M \equiv \sum_{i=1}^{N} \left[\frac{ls}{L} - \left(\frac{ds}{2L} \right)^2 \right]$$

Where N is module number. L is the connections in the network. ls is the sum of edge number and ds is sum of degree in module S.

1) 'Component' analysis method

Network portions can be called 'component' if there exist intra-contact but no inter-contact[14]. We modularized 14 microbe metabolic networks with the definition. Isolated nodes are a 'component' as well as whole networks, but it's meaningless to study on single node or whole network.

2) Girvan-Newman algorithm

Girvan-Newman (GN) algorithm is one of the most widely used clustering methods, whose clustering strategy is repeatedly identifying and removing the connections between modules. The algorithm evaluates significance of edge with edge betweenness. GN algorithm consists of four steps:1) calculate all edges betweenness; 2) remove edge with largest betweenness; 3) recalculate betweenness of the remaining edges; 4) repeat step 2) and 3). Then a hierarchical clustering tree with a top-down way is constructed by repetitive calculations.

GN algorithm uses a modularization evaluation function Q, which is the difference between real connections in module and expected connections in random graph module. Maximum integrating ΔQ in each iteration until the network only has one module. Generally, larger Q means better modularity. The time complexity of the calculation of Q is O(mn) in the network with n vertexes and m edges[1].

III. Result and Discussion

A. Modularization with 'component'

Table 2 shows that the modularized result of reconstructed metabolic network data by the definition of 'component'.

In table 2, thermophile Tma has a maximum module with 355, which also is the largest value in bacteria. The minimum number of nodes with maximal module is Sco in mesophiles, which includes 487 nodes. It represents the number of nodes in maximum module of thermophile is very different from that of mesophile. In the case of module number, the number of module in thermophilic bacteria has an average of 77.5, compared with 132.25 of the mesophilic bacteria; The average of modules number in thermophilic archaebacteria is 70.3, while the average modules in mesophilic archaebacteria is 102.3. These show that the number of modules in thermophile is significantly less than mesophile. The average ratio of modules and nodes is 14.46% in thermophilic bacteria, and the rate in mesophilic bacteria is 15.72%. The ratio of modules and nodes in therophilic archaebacteria for an average of 17.38%, compared with 20.13% of the mesophilic archaebacteria. So the community in thermophile metabolic network is tighter than that in mesophile.

TABLE II. Identification of Communities with 'Component' Analysis Methods

	Microorganism name	Maximum of module nodes	Minimum of module nodes	Number of modules	Modules/ Nodes (%)
Thermophile(Bacteria)	Aae	284	2	92	16.97
	Tel	303	2	86	16.54
	Tte	351	2	73	13.18
	Tma	355	2	59	11.15
Mesophile(Bacteria)	Ecc	505	2	135	16.05
	Sco	487	2	140	16.24
	Atc	513	1	135	17.02
	Bha	484	2	119	13.57
Thermophile(Archaebacteria)	Pab	206	2	70	17.07
	Pfu	241	2	67	15.44
	Mka	163	2	74	19.63
Mesophile(Archaebacteria)	Mac	241	1	104	22.51
	Hal	178	1	96	18.36
	Mma	259	2	107	19.53

Table 2 also shows that discrete value appears in mesophile metabolic networks but do not exist in thermophiles. And the size of maximum and minimum module is vary significantly. For example, the number of nodes in maximum module is 513 while in minimum module is 1 of Atc. And the number of nodes in the largest module has more than half the total number of nodes. A largest module nodes in Atc is 64.69% of the total number of nodes. So the modularized approach has a certain flaw.

B. Girvan-Newman (GN) clustering analysis

As the number of modules change during modularization with GN algorithm, the evaluation function Q follows a normal distribution. Best modularized result is achieved when the maximum Q value, so take the value (see Table 3) to analyze.

Q value is the standard measurement of modularized results. It has been used to analyze Zarchary karate club network[15,16]. Qian et al. evaluate the modularized result of the network consists of nodes that mapping to each software class and interface with Q value[17]. Xiong et al. combine Q value and several other algorithms to calculate the modularized results of public transportation networks[18]. Thus the measurement of modularized results with the Q value is very reliable. The Q value for thermophilic bacteria is higher than mesophilic bacteria, it means that the modular structure from the metabolic networks of thermophilic bacteria is better than mesophilic bacteria. Nevertheless, there is almost no difference in archaebacterial. Tma with lowest Q value also reaches 0.88, therefore GN algorithm has a good effect on modularize.

TABLE III. Identification of communities with Girvan-Newman algorithm

	Mirco organism name	**Q value**	**Number of modules**	**Number of maximum subset nodes**	**Ratio of modules and nodes（%）**
Thermophile(Bacterium)	Aae	0.90	54	298	9.96
	Tma	0.88	72	54	13.85
	Tel	0.90	98	48	17.69
	Tte	0.90	91	29	17.20
Mesophile(Bacterium)	Ecc	0.89	161	45	19.14
	Sco	0.89	161	85	18.68
	Atc	0.90	151	79	19.04
	Bha	0.89	139	76	15.85
Thermophile(Archaebacteria)	Pab	0.91	84	33	20.49
	Pfu	0.91	78	30	17.97
	Mka	0.93	81	29	21.49
Mesophile(Archaebacteria)	Hal	0.92	103	33	22.29
	Mac	0.92	113	36	21.61
	Mma	0.91	120	67	21.90

The average number of modules in mesophilic bacteria is 153, it is almost twice as much as in thermophilic bacteria, which is 78.75. The average number of modules in thermophilic archaebacteria and mesophilic archaebacteria are 81 and 112 respectively. The ratio of modules and nodes in different kinds of organisms are 14.68% (thermophilic bacteria), 18.18% (mosophilic bacteria), 20.00% (thermophilic archaebacterial) and 21.93% (mesophilic archaebacterial). Whether the number of modules or the ratio of modules and nodes in thermophiles is obviously less than that in mesophiles. Hence the overall modularized result of thermophiles is better than mesophiles. But Aae is a special case in Table 3, whose number of nodes in maximum subset is much greater than the total number of nodes itself. And the nodes in the maximum subset are all discrete.

IV. 4. CONCLUSION

In this paper, 14 kinds of metabolic network are modularized in terms of the definition of ‘component’ and GN algorithm. The ratio of number of modules and total number of nodes in thermophiles and mesophiles are 15.71% and 17.61% respectively with the modularized result by the definition of ‘component’. Meanwhile, the ratio of number of modules and total number of nodes in thermophiles and mesophiles are 16.90% and 19.79% with Girvan-Newman algorithm. Thermophiles are more modularity than mesophiles, which indicates that the topological structure, especially inner modular structure of metabolic networks is influenced by environment temperature. Compared with the modularized result that the node distribution is uneven by the definition of ‘component’, GN algorithm is much better.

REFERENCES

[1] Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences USA, 2002, 99(12): 7821-7826.

[2] Albert R, Barabási A-L. Statistical mechanics of complex networks. Reviews of Modern Physics 2002, 74(1): 47-97.

[3] Barabási A-L, Albert R. Emergence of Scaling in Random Networks. Science, 1999, 286(5439): 509-512.

[4] Guimera R, Nunes Amaral LA. Functional cartography of complex metabolic networks. Nature, 2005, 433(7028): 895-900.

[5] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818.

[6] Tian Ye, Liu Dayou, Yang Bo. Application of Complex Networks Clustering Algorithm in Biological Networks. Journal of Frontiers of Computer Science and Technology, 2010, 4(4): 330-337.

[7] Ding Dewu,Lu Kezhong, Xu Wenbo, Wu Pu, Huang Haisheng. Functional Modules in B. thuringiensisMetabolic Network Based on SAA. Computer Engineering, 2010(13): 162-3+6.

[8] Li Junjin, Xiang Yang, Niu Peng, Liu liming, Lu Yingming. New complex network clustering algorithm. Application Research of computers, 2010,27(6): 2097-2099.

[9] Flake GW, Lawrence S, Giles CL, et al. Self-organization and identification of Web communities. Computer, 2002, 35(3): 66-70.

[10] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113.

[11] Guo XingLi, Gao Lin, Chen Xin. Models and Algorithms for Alignment of Biological Networks. Journal of Software, 2010, 21(9): 2089-2106.

[12] Zhang S, Wang R-S, Zhang X-S. Identification of overlapping community structure in complex networks using fuzzy -means

clustering. Physica A: Statistical Mechanics and its Applications 2007, 374(1): 483-490.
[13] Ma H, Zeng A-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 2003, 19(2): 270-277.
[14] Liu Jun. Lectures on Whole Network Approach, A Practical Guide to UCINET. 2009, Due press , Shanghai.
[15] Zhu Xiaohu, Song Wenjun, Wang Chongjun, Xie Jinyuan. Improved Algorithm Based on Girvan-Newman Algorithm for Community Detection. Journal of Frontiers of Computer Science and Technology, 2010, 4(12): 1101-1108.
[16] Liu Shaohai, Liu qingkun, Xie fuding, An na. Algorithm for detecting community structures in complex networks. Computer Engineering and Design, 2009, 30(20): 4708-10+14.
[17] Qian Guanqun, Zhang Lin, Zhang Li. Two-phase software clustering method based on complex network theory. Journal of Bei jing University of Aeronautics and Astronautics. 2009, 35(12): 1438-1442.
[18] 18. Xiong Yan, Wu Bin, Du Nan, Ye Qi, Pei Xin. A Hierarchical Clustering Algorithm of Public Traffic Network Based on Maximal Cliques. Journal of Computer Research and Development, 2007, 44(z2): 123-128.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Performance Analysis for Fast Parallel Recomputing Algorithm under DTA

Wanfeng Dou
Dept. Computer Science & Technology
Nanjing Normal University
Nanjing 210023, China
e-mail: douwanfeng@njnu.edu.cn

Shoushuai Miao
Dept. Computer Science & Technology
Nanjing Normal University
Nanjing 210023, China
e-mail: ms_shuai@126.com

***Abstract*—With the rapid increasing of spatial data resolution, the huge volume of datasets makes the geo-computation more time-consuming especially in operating some complex algorithms. Parallel computing is regarded as an efficient solution by utilizing more computing resource. The stable and credible services play an irreplaceable role in parallel computing, especially when an error occurs in the large-scale science computing. In this paper, a master/slave approach of implementing the fast parallel recomputing is proposed based on redundancy mechanism. Once some errors in application layer are detected, the original data block with computation errors is further partitioned into several sub-blocks which are recomputed by the surviving processes concurrently to improve the efficiency of failure recovery. The multi-thread strategy in main process is adopted to distribute data block, detect errors and start recomputing procedure concurrently. The experimental results show that the proposed method can achieve better performance efficiency with fewer additional overhead.**

Keywords- parallel computing; digital terrain analysis; fault tolerance; fast parallel recomputing

I. INTRODUCTION

Digital Terrain Analysis (DTA) is a digital information processing technology of computation of the terrain attributes and feature extraction on the basis of the Digital Elevation Mode (DEM). With increased precision and accuracy, DEMs have gone from 1,000 meter resolutions 5-10 years ago to 1-5 meter resolutions today in many areas. As a result of the increased precision and file sizes, many land surface parameters, such as slope, profile curvature and hydrologic land-surface parameters for lower resolutions and smaller DEMs become prohibitively time-consuming when being applied to high-resolution and large volume of data. Hence, the parallel computing has becomes a fundamental tool for geographic information science [1-2].

Large-scale scientific computing brings more computation tasks so that the running time is longer than the applications in the small scale. Meanwhile, the failure risk become higher and it could cause a lot of resource waste including the time and storage. The checkpointing technique [3] is widely used for fault tolerance. The basic idea is that the current state and data should be saved to stable storage when processing the key position of a process. A large number of the checkpoints will be stored in the parallel system periodically. Whenever a process fails, all processes related to this process have to be rolled back to the last checkpoint to restart the computation. The operation of the checkpoints will lead to the huge volume of data transmission through I/O, which becomes a performance bottleneck.

To speed up the recovery procedure, Yang et al. [4] put forward a new application-level fault-tolerant scheme based on parallel recomputing, called Fault-Tolerant Parallel Algorithm (FTPA). However, the method is mainly applied to instruction level of a program that will affect whether or not the parallel recomputing is done. When the failure part of the program cannot be repartitioned again, then parallel recomputing could not be finished, so the performance of recomputing will not be provided. Moreover, Yang et al also proposed an improved approach that can solve program division and workload redistribution [5]. From the above, these approaches discussed the key issues about the program partitioning rather than data partitioning. A parallel recomputing method for parallel digital terrain analysis is first presented by Miao and Dou to improve the efficiency with redundancy computing mechanism [6]. They presented an improved error detection scheme which adopts the multiple threads strategy to check whether or not the errors exist in the computation of a data block so as to reduce the overheads of error detection [7]. Song et al implemented a fault-tolerance parallel computing algorithm adopting checkpointing technique for digital terrain analysis [8].

II. FRAMEWORK OF FAST PARALLEL RECOMPUTING

The fast parallel recomputing is based on redundancy mechanism. Its basic principle is that the task with a data block and the task with its copy are first executed on two different nodes concurrently. After these two tasks are finished, the results of them are sent to the process which is run on master node and compared whether these two results are consistent or not. Once an error will occurs when the results are not consistent. When the master node receives this error information of this data block, then it will partition the data block into a group of small sub-blocks, for instance four sub-blocks. Then these sub-blocks are distributed to the surviving processes to compute in parallel. Finally, the results of these sub-blocks are sent to the master node to fuse as the result of the original data block. The whole process of fast parallel recomputing consists of five steps:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.19

Step 1, Redundancy computing. In normal parallel computing, in order to detect whether or not there is a computation error in a process with the data block, redundancy computing strategy is usually adopted. In dual-redundancy computing, two processes, P and P', with the data block B and its copy B' are executed concurrently so that their results, R and R', are used to compare each other to justify the correctness.

Step 2, Error detecting. When the computation results of two processes, R and R', are received by the master node, they are used to compare whether or not they are consistent. If they are consistent, one of the results is saved. If the results are not consistent, the original data block is repartitioned into several small sub-blocks so as to recompute them to remedy the error.

Step 3, Data repartitioning. Once a computation error is detected, the data block with computing errors will be repartitioned into 4 or 16 sub-blocks according to system's computation resources.

Step 4, Parallel recomputing. Several processes are started execute these sub-blocks repartitioned. These processes executes in parallel so that the result is fast recovered using fewer time.

Step 5, Results fusing. The results of sub-blocks are sent to the master node and are fused into the result of the original data block.

III. IMPLEMENTATION OF FAST PARALLEL RECOMPUTING IN MASTER/SLAVE MODE

Designing a fast parallel recomputing is to incorporate the parallel error remedying scheme in parallel digital terrain analysis. The major characteristic of the fast parallel recomputing is that the data block of a failed process is repartitioned into some small sub-blocks by master node and then these sub-blocks are redistributed to and recomputed by the idle processes concurrently to gain the right result of the original data block.

A. Data distribution strategy in single process mode

In order to implement the fast parallel recomputing procedure, we adopt a master/slave structure in which the master node is responsible for data distribution, error detection and results fusion. The slave node is responsible for the computation of each data block. It is an easy approach to implement the scheduling tasks in regular sequence, so we first discuss a data distribution strategy in sequence.

Fig.1 shows the execution time of scheduling tasks in sequential. Firstly, the main process on master node reads the whole data from the disk and distributes each data block to two computing processes in different slave nodes, in which one is a normal process and the other is a copy of this process. Hence, there are 2n computing processes. Secondly, all 2n processes will execute the computing tasks once they receive the required data block. The main process is possible to wait a while after finishing the distribution of all data blocks in which time the first computing process could not finish the computing task. It depends on the time of each computing process executing a task in slave node and data distribution time in main process. If $2nt_d-1>t_c$, then the first computing process is required to wait for sending the result to the master process until the master finishes the data distribution, otherwise, the master process is required to wait for receiving the computing result until the first computing process finishes the computing task.

As seen in Fig.1, once the master process receives the computing results of the normal process and its copy process, it will check whether the two results are consistent by comparing every value in the results. If the results are consistent, it means that the computing procedure is right and no error exists, the master process will receive next result of subsequent process. Otherwise, the result is not correct, the master process must repartition this block into 4 sub-blocks and distribute to 4 idle computing processes to recomputing. Thirdly, the master process receives the recomputing results and fuses them into the result of the original data block. Finally, the master writes the whole result to the disk.

Suppose that T_{read} is the time of reading the whole data to memory, t_d is the time of distribution of each data block, t_c is the computing time of each data block, t_r is the time of receiving the computing result of each data block, t_{cp} is the time of comparing two results, and T_{write} is the time of writing the result back to the disk.

Due to the reading data time is long which results in the long waiting for the computing process, so we may improve the way of reading and distributing data. After a data block is read into memory the master process may distribute it at once.

B. Data distribution strategy in multi-thread mode

In order to provide the high efficiency in data distribution and error detection, we adopt the multi-thread technology based on shared memory mode. Each thread is responsible for data distribution and error detection. When there is any error after comparing, the thread is also responsible for redistributing the logic data block on which is occurred computing error to a new process to execute recomputing.

The main process in master node reads the whole data into main memory from the disk and creates all threads according to the number of partitioned data blocks. Each thread is responsible for distributing each data block partitioned in advance to the process and its copy process in slave nodes. After finishing the data block distribution each thread is waiting for receiving the computing results of each data block. Once the results of the data block and its copy are received, the thread compares the whether the results are consistent or not. If the results are consistent, it shows that the computation is correct and the thread will submit the result to memory to fuse. Otherwise the result is not correct, the respective data block is repartitioned into 4 sub-blocks and redistributed to 4 new processes to execute the recomputing in parallel. Only when the results of all data

blocks are gained and detected to be correct, and then fused to whole result set the thread could be closed. Finally, the main process saves and writes the whole result set to disk file.

The process in a slave node and its copy process in another slave node are responsible for receiving a data block, computing the result and sending the result back to respective thread. The process receives the data block and starts to compute the results of data block. Finally, it sends the result of data block to the thread according to the appointment in advance.

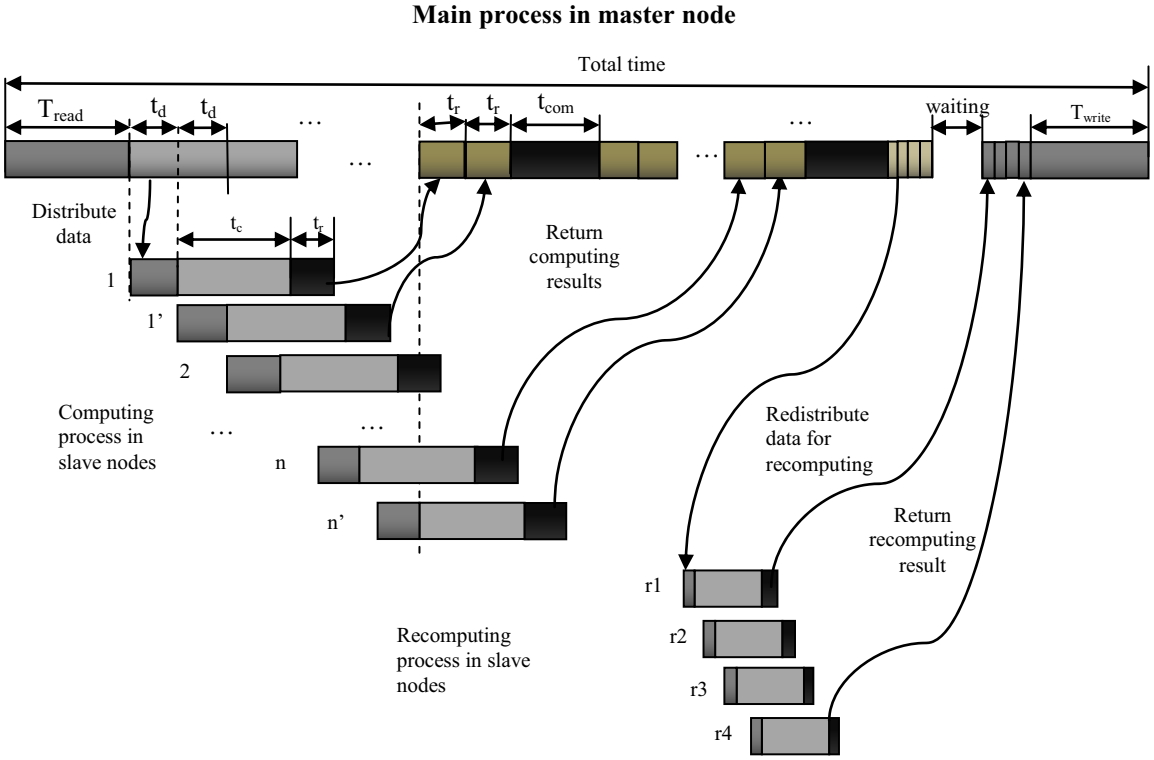


Figure 1 Model of data scheduling in mater/slave mode

IV. Experiments and Analysis

The experiments were performed on a small scale cluster system. Each node is equipped with an Intel XeonE5645, 2.8 GHz with quad-core processors and 8GB memory. The nodes adopt the Gigabit Ethernet connectivity. The master-slave parallel computing model is adopted. A primary node is responsible for distributing data and recycling results. The software environments have the GDAL 1.6.1, OpenMP 1.5.4, MPICH2. Two kinds of datasets are employed as the testing data, the size of smaller dataset is 1.61GB and the one of bigger dataset is 6.9GB. The data type is floating-point and the type of image is TIFF format.

In this paper, we implemented a parallel recomputing algorithm with mater/slave mode in which the master node is responsible for data distribution, results comparison and result fusion and saving to disk, and slave node is responsible for computing task with a data block.

We adopted dual redundancy mechanism in which the master process distributes data block two times and one is distributed to a process and another one is to its copy process. Fig.2 shows the results with two DEM datasets, one is the size of 1.6GB and another one is the size of 6.9GB. As shown in Fig.2, we summarize that the total computation time declines with the increase of number of process (data blocks). Since only a failure of process occurs only four new processes are started for executing the recomputing procedure, hence the overheads of the system are small. But only a main process in master node is responsible for data distribution, result comparison and result fusion and etc. and is so busy that the computation processes from slave nodes have to wait for communication. So the total computation time does not decline quickly and when the number of processes which are responsible for the computation attains to a certain value the total time will increase slowly with the number of processes or data blocks.

The above experiments permit the master process in master node to distribute data in sequence, receive and compare rightness of two results from the computing process and its copy process respectively. The advantage is that the data is read into memory from disk orderly which does not incur I/O contention and the disadvantage

is that the processes in slave node will wait for data distribution. The following experiments adopt multi-threads may in master process to distribute data blocks in parallel and compare the correctness of results and determine whether or not the recomputing procedure is started.

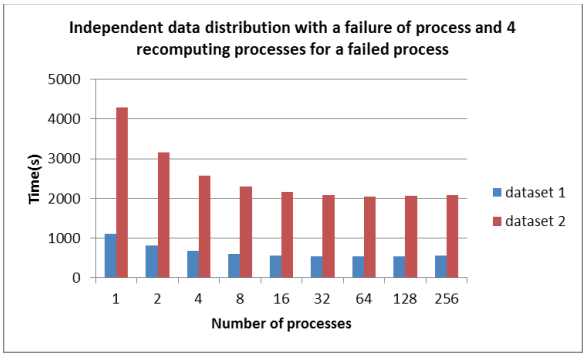


Figure 2. Total computation time with the number of process under independent data distribution

Fig.3 shows the total computation time with the number of thread in main process. For a failed process 4 processes will be started to execute the recomputing procedure. As shown in Fig.3, the time declines with the increase of thread number for two datasets. The reason is that there are a lot of threads to deal with the distribution of every data block, results comparison and starting recomputing procedure in parallel. The parallel data distribution using multithread technique can provide high efficiency in data distribution and error detection comparing with independent data distribution.

Obviously, the total computation time in parallel data distribution is lower than that one in independent data distribution. Furthermore, the time in parallel data distribution declines quickly with the number of thread comparing with independent data distribution.

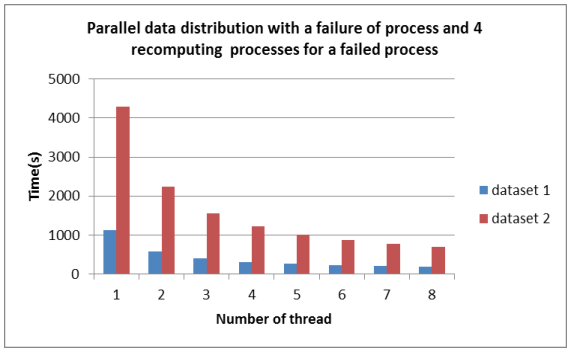


Figure 3. Total computation time with the number of thread under parallel data distribution

V. CONCLUSIONS

In this paper, a fast parallel recomputing algorithm for parallel digital terrain analysis is implemented to achieve a fast self-recovery of error results. It is time-consuming to deal with the fault-tolerant parallel computing for the whole of DEM data because more time is required to detect and correct errors. Hence, the data partition strategy and corresponding multi-thread scheme are adopted to solve these problems. The computing time will be shortened and the overall performance of parallel computing will be optimized by repartitioning the data block with errors into several smaller sub-blocks. Moreover, the performances of fast parallel recomputing are evaluated in different data size with the slope algorithm on a cluster system. The experimental results show that the overhead of fast parallel recomputing does not result in more overheads of the system when a failure occurs.

However, our work addresses in the case that there is no dependent relationship between data blocks. In the future, our efforts will mainly focus on two aspects. Firstly, the performance of fast parallel recomputing will be evaluated in different digital terrain analysis algorithms. Secondly, further work will be focused on the fault tolerance with the dependent data blocks. The I/O performance is also further needed to consider with an introduction of redundancy computing.

ACKNOWLEDGMENT

This work has been substantially supported by the National Natural Science Foundation of China (NO. 41171298)

REFERENCES

[1] G. Chen, G. Sun, and Y. Xu “Integrated research of parallel computing: Status and future”. J. Chinese Sci Bull, Vol.54, pp.1845-1853, Nov., 2009.

[2] G. Lecca, M. Pditdidier, and L. Hluchy “Grid computing technology for hydrological applications,” J. Hydrology, Vol.403, pp.186-199, Jan., 2011.

[3] J. Plank, K. Li, and M.A. Puening “Diskless checkpointing,” IEEE Trans. Parallel Distrib. Syst., Vol.9, pp.972-986, Oct., 1998.

[4] X. Yang, Y. Du, P. Wang, and et al. “The Fault tolerant parallel algorithm: the parallel recomputing based failure recovery,” Int. Con. on Parallel Architecture and Compilation Techniques, IEEE Press, 2007, Brasov, pp.199-209.

[5] H. Fu, Y. Ding, and W. Song, “An application level checkpointing based on extended data flow analysis for OpenMP programs,” J. Chinese Journal of Computers, Vol.32, pp.38-53, Oct., 2010.

[6] S. Miao, W. Dou, and Y. Li “Research on the fast parallel recomputing for parallel digital terrain analysis,” W. Xu et al. (Eds.): GRMSE 2014, CCIS 482, Springer-Verlag Berlin Heidelberg, 2014, pp.244-251.

[7] S. Miao, W. Dou, and Y. Li “Study on error-detecting approach for fault tolerance recomputing oriented parallel digital terrain analysis,” Proc. of on DCABES, Xianning, China. Eds: Craig Douglas and Guo Yucheng, IEEE Press, 2014, pp.148-151.

[8] X. Song, W. Dou, G. Tang, and et al. “A diskless checkpointing algorithm for cluster architectures applied to geospatial raster data processing,” Journal of Algorithms & Computational Technology, Vol.8, pp.369-387, Dec., 2014.

Use Pre-record Algorithm to Improve Process Migration Efficiency

Shan Zhongyuan
Northeastern University
College of Information Science and Engineering
Shenyang, China
Shan52@163.com

Qiao Jianzhong
Northeastern University
College of Information Science and Engineering
Shenyang, China
qiaojianzhong@mail.neu.edu.cn

Lin Shukuan
Northeastern University
College of Information Science and Engineering
Shenyang, China
linshukuan@mail.neu.edu.cn

Zhang Qiang
Northeastern University
College of Information Science and Engineering
Shenyang, China
zqzq53373931@163.com

***Abstract*—Process migration is one of the most important features in parallel and distributed computing. It enables dynamic load balance and makes better utilization of computing resource. Post-copy is a very efficient migration algorithm but it needs process to resume on destination node with incomplete address space which may significantly reduce its efficiency especially at the initial phase. To solve this problem, we propose a new algorithm named as Pre-record. It will prolong process execution on host node for a short while before suspend and record the used memory pages. While transmitting process address space, these recorded pages will be transferred preferentially. So, At the initial phase of process resume on destination node, the needed memory pages have already been stored, no pages faults occurs. We evaluate Pre-record performance through simulation test, make a comparison with the other algorithms, and the results shows that Pre-Record could significantly reduce page faults number and improve process migration efficiency.**

Keywords- address space; distributed computing; page fault; post-copy; process migration

I. INTRODUCTION

Process migration could transfer the running process among a group of computers. Using this technology, distributed system could easily achieve dynamic load balance, reduce communication expense, enhance fault resilience and so on. It has already been one of the key features of distributed system [3]. Various clusters, grid and cloud computing models all choose suitable process migration algorithm to improve system efficiency such as MOSIX, Sprite, the European data grid (Data Grid), Microsoft System Center and so on.

There are several process migration algorithms, such as Eager-copy, Lazy-copy, Pre-copy and Post-copy. All the algorithms have their own advantages, but also have short-comings. Generally speaking, migration algorithm should shorten process freezing time to ensure user transparency, should reduce the dependence on the host node and ensure the execution efficiency after process resume running. Eager-copy is the earliest migration algorithm [7], and has been imple-mented in lots of systems like Charlotte [5], and Amoeba [6]. The migrated process could run properly on the destination node. But it is criticized due to long freezing time caused by transmitting the entire address space. Lazy-copy has the shortest freezing time, but with the worst cost of execution efficiency after process resumes. It also has a high level residual dependency on the host node. Lazy-copy is mentioned as copy-on-reference in Edward R. Zayaz's literature [4]. Pre-copy algorithm has a very short freezing time and can ensure the execution efficiency after migration. But It has to face to the convergence problem while migrate the high frequency read-write memory process, and parts of address space need to transfer multiple times which may increase network traffic. Pre-copy also has residual dependency problem.

Post-copy is the most widely discussed algorithm in recent years [1] [2] [8]. It reduces the freeze time by resume running just after kernel data transmitted, transmit the entire address space to avoid residual dependency, and use parallel approach to acquire memory pages. But, the process needs to resume with incomplete address space, so it also suffer the pain of page faults, as the same as Lazy-copy.

To improve the shortcomings of the existing migration algorithms, we propose a new algorithm named as Pre-record. It would prolong process execution on host node and record the visited memory pages. Through this way, it could predict the memory pages which will be visited on the destination node while the process resumes. These recorded pages will be migrated preferentially before other address space. So, the number of page faults could be significantly reduced. Pre-record also retains the advantages of short freezing time and little residual dependency on host node.

II. PRE-RECORD ALGORITHM

A. Principle

The efficiency of migration algorithm is closely tied with the following three points: less freezing time, less page faults and none residual dependency. Existed migration algorithms have their own advantages, but hardly to cover all of the three points.

Freezing time means process stop execution time during migration. If the freezing time is too long, it may depress user experience. Load of host node can be reduced only after the process is completely separated from host node. In some cases, host node may be unworkable due to unrecoverable error or overload. In such kind of situation, the host node has no way to join in further execution. So reduce the dependence on host node during migration which is called residual dependency [3] is quite important for an excellent

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.20

algorithm. Migration algorithm also must ensure execution efficiency after process resume running on destination node.

Consider all above, we proposed Pre-Record. Pre-record uses the time difference of process execution on host node and destination node, transfers the forecast work of needed memory pages to the destination node. Pre-record prolong the process execution on host node, until the transition of kernel address space is finished and then stops the process. During this period of execution, process will visit parts of the address space, memory pages, stack, files etc. Because the process's execution on destination node is a continuation of the host node, so when the process resume running on destination node, it will visit the same part of address space immediately. Host node records these memory pages and transmits these recorded pages right after the kernel address space. In this way, the needed memory pages have been already transmitted when the process resumes on destination node. But the results of this period of execution will not be inherited by the destination node, and the dirty pages will not be transmitted either.

B. Pre-record the memory pages

In order to shorten the freezing time, only transmit the kernel data before resume the process on destination node is used most regularly. Running with incomplete address space, the process may be suspended frequently because of the page faults. And it may seriously reduce the execution efficiency. The main idea to solve the problem is to predict the next memory page which will be visited, and ask host node to transmit it before it is needed.

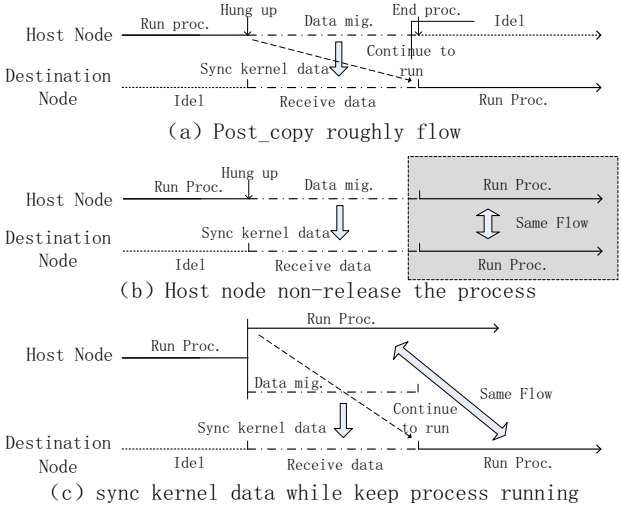


Figure 1. Schematic diagram of memory pages pre-record algorism

Consider that the execution of process on destination node will inherit the result on host node. We can do some job to predict which memory page will be visit on host node. Take Post-copy as an example, the process is hung up on host node firstly, then through the synchronization of kernel address space, it resumes running on destination node, as shown in Figue1(a). If the host node does not hang up the process after transferring the kernel address space, but continue executing the process, then the host and destination nodes will execute the same process and visit the same memory page, as shown in Figure 1(b). In Post-copy, the process is hung up on host node while transmitting kernel address space. In fact, this period could be utilized to keep process running and record the visited address space, such as memory pages, stacks, files etc. When the process resumes on destination node, it will visit the same parts of address space, as shown in Figure 1(c). So the predict job has been delivered to host node. Then we transfer the recorded memory pages at first. In this way, the requested memory pages have already been transferred while the process resume running on destination node. So there will not be page faults during this period.

C. The procedure of Pre-record algorithm

Pre-record algorithm's work flow could be sketched as following (see Figure 2.). we choose to hung up the process to stop the pre-record phase when the kernel address space is migrated.

(1)Host and destination node come to an agreement to migrate the process away.

(2)Host node starts to transmit the kernel address space. (minimal necessary subset of the process state) to destination node.

(3)At the mean while, host keeps executing the process.

(4)The visited memory pages are recorded into the table named as DirtyPage_Map.

(5)When the migration of kernel address space is completed, host node hangs up the process.

(6)Host node starts to transfer the recorded memory pages stored in DirtyPage_Map to destination node.

(7)The process's kernel address space is rebuilt, and it resumes executing on destination node.

(8)Host node transmits rest address space, in parallel with the execution on destination node.

(9)Page fault happens, destination node send page request to host node.

(10)Source node interrupts the sending sequence, and answers the page request priority.

(11)After the entire address space completely transmitted, the entire process will be deleted from source node. And the migration is complete.

The execution of process on host node is divided into two phases: normal execution and pre-record execution. Before starting to migrate the address space, host node has to establish communication with destination node, and destination node gets ready to receive the process. At this period, the process is in normal execution phase. When the prepare work finished, host node starts to transmit the kernel address space and the process execution switch into the pre-record execution phase. In pre-record execution phase, host node will record the visited memory pages and store them in to the preinstalled table DirtyPage_Map. When the process resumes running on the destination node, it will inherit the normal execution results to continue, repeat the pre-record execution phase again, so it needs to visit the memory pages which have already been recorded in DirtyPage_Map table. Give priority to the memory pages in DirtyPage_Map table while transmitting the process address space, can ensure that

the process will not be suspend because of page fault while it repeats the pre-record execution phase on destination node.

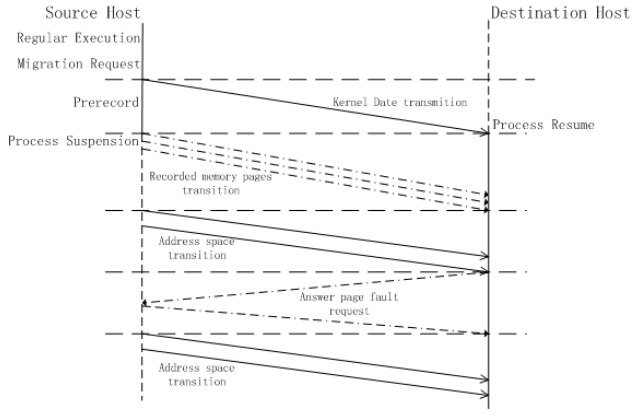


Figure 2. Pre-record algorithm working flow

The migration of process address space is made up of four phases: (1) Migrate the kernel address space. In order to make the process resume running at the first time, Pre-record transfers the kernel address space which are necessary for process running, and then resume the process on destination node immediately. Lazy copy and the Post-copy algorithm also uses a similar approach. (2) Migrate the pre-recorded memory pages. After the kernel address space, host node will continue to transfer the recorded memory pages which are stored in DirtyPage_Map. (3) Migrate the remains. In order to free itself, host node continues to transfers all the remained address space after the transition of the pre-recorded pages. (4) Migrate according to needs. While the process running on the destination node, it may be suspended because of the page faults. At this time, destination node will ask host node for the data. Host node constantly monitors such kind of page requests, answers it with highest priority, interrupts the current sending queue, and sends the requested pages first.

D. Group transfer

The locality principle ensures that once a process starts working with a group of pages, it sticks with them without addressing other pages for quite a while. So, after the process resumes on the destination host, if the process works with a memory page at one time, it is likely to visit the adjacent pages in the next period. Similarly, when process runs on the destination node, and be suspended by page fault, the memory pages adjacent the missing page also have a very high chance to be visited later. So, when page fault happens, the source node does not just send the requested page but also send the nearby ones. In this way, we could greatly reduce the incidence of potential page fault over a period of time. According to this principle, we proposed the predict memory algorithm which is adapted to the Pre-record algorithm.

While transferring kernel data, source host continue executing the process and record all the memory pages visited by the process. Pre-record algorithm maintains an array T to record all the pages' addresses. There is also another array S to record the time when the page is visited. For simplicity, we store the intervals between the two pages visit in array S. For example, the process visited p_1, p_2, p_3 at 0.5s, 0.7s, 1.1s. So we have array T= {p_1, p_2, p_3}, and array S= {0, 0.2s, 0.4s}.

The total size of the kernel data is defined as M, and the time of transferring this part is defined as T. The migration rate v on source node is v=M/T. We assume destination node has the similar performance as source node. And T = {p_1, p_2, p_3,…, p_i}, S = {t_0, t_1, t_2, … , t_i}. The size of memory frame is m. After process visited pi and before visit p_i+1 on destination node, the number of memory page sent from source node is record as W_i. So we have:

$$W_i = (v * t_i) / m = (M * t_i) / (T * m) \qquad (1)$$

After the migration of kernel data is finished. p_1 is transferred to destination node at first. And then transfer p_2 and the following W_2 frames. The third time is W_3 and the k time is W_k ($0 \leqslant k \leqslant i$).

At the k time, the number of memory pages could be sent by source node is W_k. During sending these W_k pages, source node finds that one of these pages may have already been sent before. According to the memory predict algorithm, the following pages must have already been transferred, too. So, this round of sending could be broke off and start the next round of sending.

III. COMPERATION

Analyze the efficiency of process migration algorithm and make a compare is a very complex problem. There are lots of situation needs to consider, such as the performance of the test machine, kind of process, network traffic situation, the size of address space, and so on. Considered about the Comparison method mentioned in [13], we will compare Pre-record algorithm with the others in the following four points: (1)The freeze time between processes suspend on source node and resume on destination node. (2)The interrupted frequency while process is running on destination node. (3)The effect on network traffic. (4)Residual Dependency.

The advantages and disadvantages of migration algorithms are shown in the table 1.

TABLE I. THE COMPARISON OF MIGRATION ALGORITHMS

Algorithm	Freeze Time	Suspend Times	Network Traffic	Residual Dependencies
Eager-copy	High	Especially Low	Medium	No
Lazy-copy	Low	High	Low	Yes
Pre-copy	Low	Especially Low	High	No
Post-copy	Low	Medium	Medium	No
Pre-record	Low	Low	Medium	No

Pre-record resumes process on destination host after the transfer of kernel data, so process could resume ASAP. The delay time could be as short as lazy-copy and shorter than any other algorithm. Pre-record will transfer the necessary memory pages at first, and also transfer the nearby pages depending on the locality principle. It goes further than the post-copy on reducing page faults. Transfer address space may increase network traffic load. Lazy-copy with partly

migration has less network load. Pre-record uses full migration. Entire address space will be transferred. Pre-copy does not only transfer the entire address space but also transfer the dirty pages. It will increase network load additionally. Pre-record also does not have residual dependency problem.

IV. EXPERIMENTS

We use Microsoft Visual C++ write test software to simulate the procedure of process migration, and carry out experiments to compare the efficiency of each migration algorithm.

(1) The compare of the page faults distribution of Post-copy and Pre-record algorithms

We simulated to run Post-copy and Pre-record algorithm, and recorded the address of the missing page when page fault happens. The result is in Figure 3.

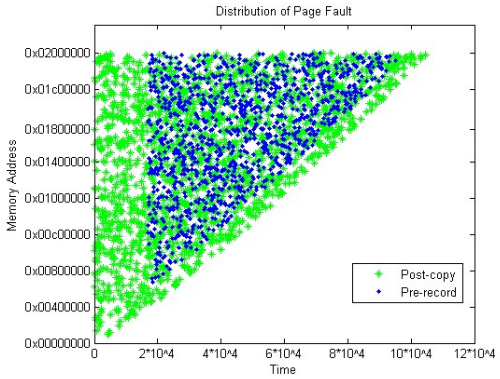


Figure 3. The distribution of page faults of Post-copy and Pre-record

As Figure3, it has the highest the number of page faults at the initial phase of process resumes running on destination node with Post-copy. In the experiment of Pre-record, there is no page fault while the process repeats the pre-record execution phase of host node, and the overall number of page faults is significantly reduced, the migration process will also be completed in a shorter period of time.

(2) The influence of network transmission speed on the efficiency of the Pre-record algorithm.

In the experiment, we set the process to visit address space randomly and set the migration working under different network delay. Observe the efficiency of Pre-record algorithm under different network conditions, and make a compare with Post-copy algorithm. The result is in Figure 4.

Experiment result shows that Pre-record has better efficiency than Post-copy under different network situation. From the test results it can be seen that network situation may greatly impact on the migration algorithm efficiency. For the same process, the better network traffic speed is, the higher migration efficiency will be; while under the same the network traffic situation, the process which read the memory less frequently has the better migrate efficiency. And the key reason is the relationship between network transmission delay (SendTime) and the frequency of process visiting memory pages(SwitchTime). When SendTime is less than SwitchTime , Pre-record efficiency is better than Post-copy undoubtedly. When SendTime is equal with or larger than SwitchTime, host node could not transfer the memory pages before needed, so the efficiency of Pre-record will decrease down to the same as Post-copy.

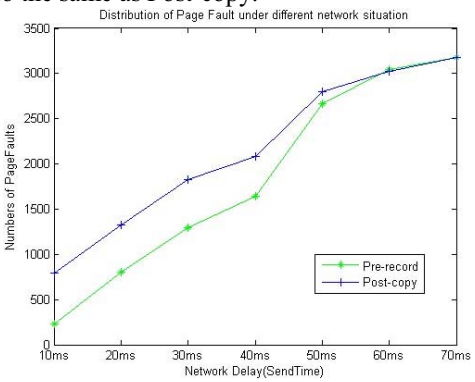


Figure 4. The page faults of Post-copy and Pre-record under different network delay.

V. CONCLUTION

This paper presents a new process migration algorithm named as Pre-record algorithm. Pre-record has less freezing time by just transferring kernel address space before process resumes, and successfully eliminate residual dependence by pushing all address space at background. Pre-record predicts memory pages on host node and uses group transfer to reduce pages faults then to improve process execution on destination node. The comparative analysis and simulation experiments show that the Pre-record has higher migration efficiency. Pre-record is worth for further research and would play a greater role in parallel and distributed computing in the future.

REFERENCES

[1] Roy S.C.Ho, Cho-Li Wang and Francis C.M Lau, "Lightweight Process Migration and Memory Prefetching in openMosix," IEEE International Parallel and Distributed Processing Symposium, Miami, 2008, pp. 1-12.

[2] Michael R.Hines and Kartik Gopalan, "Post-copy Based live Virtual Machine Migration Using Adaptive Pre-paging and Dynamic Self-Ballooning." ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, Washington, 2009, pp. 51-60.

[3] Dejan S.Milojicic, Fred Douglis, Yves Paindaveine, Richard Wheeler, and Songnian Zhou, "Process migration," ACM Computing Survey, Vol.32, No.3, pp. 241–299, September 2000.

[4] Edward R.Zayas, "Attacking the Process Migration Bottleneck," Proceedings of The 11th ACM Symposium on Operating Systems Principles. Austin, November 1987, pp. 13-24.

[5] Yeshayahu Artsy, Raphael Finkel, "Designing a process migration facility: the Charlotte experience," IEEE Computer, September, 1989, Vol.22, No.9, pp. 47-56.

[6] Chris Steketee, Wei Ping Zhu, Philip Moseley, "Implementation of process migration in Amoeba," 14th International Conference on Distributed Computing Systems Principles, Poznan,1994, pp. 194-201.

[7] Michael L.Powell, Barton P.Miller, "Process Migration in DEMOS/MP," ACM SIGOPS Operating Systems Review, 1983, Vol.17, No. 5, pp. 110-119.

[8] Ellard T.Roush and Roy H. Campbell, "Fast dynamic process migration," Proceedings of the 16th International Conference on Distributed Computing Systems, 1996, pp. 637-645.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Parallel Algorithm Study of Petri net Based on Multi-core Clusters

Wenjing LI
School of Logistics Management and Engineering
Guangxi Teachers Education University
Nanning, China
e-mail:liwj@gxtc.edu.cn

Zhong-ming Lin, Ying PAN ,Ze-yu Tang
College of Computer and Information Engineering
Guangxi Teachers Education University
Nanning, China

***Abstract*—The parallel algorithm of Petri net based on multi-core clusters is put forward in order to make the Petri net system with concurrent synchronous function realize parallel control and running. First, select different Petri net structures and conduct transformation, and give the partitioning method of the subnets of place invariant-based Petri net system. Then, put forward the parallel algorithm of Petri net based on multi-core clusters according to the MPI+OpenMP+STM (STM, Software Transactional Memory and transactional memory) three-level parallel programming model and combining with the parallelized analysis of the changes of internal subnets and among the subnets. The experiment results show that the algorithm can better reflect the actual running process of Petri net system, and it is a feasible and effective method of realizing the parallel control and running of Petri net system.**

Keywords- Multicore clusters; Petri net; Petri net structure and transformation; Subnet partitioning; MPI+OpenMP+STM parallel model; Parallel algorithm

I. INTRODUCTION

Petri net is the most direct, natural and accurate indication of concurrence, mutual exclusion, synchronization during the running of a complex system, and it has very important significance in its parallelization and algorithm design and applying it into the simulating execution and operation of actual Petri net. The parallelizing process of Petri net includes five steps: first, conduct extraction and transformation to the Petri net structure; second, solve the P/T net place invariant and conduct subnet partitioning; third, conduct parallelized analysis to the changes of the internal subnets and among the subnets; fourth, design different parallel algorithms combining with different parallel models; fifth, implement and verify the algorithms.

II. EXTRACTION AND TRANSFORMATION OF PETRI NET STRUCTURE MODEL

Use different Petri net models to describe the structure and behavior of the system for different applications or solving different practical problems, so there are many different kinds of Petri net models while these kinds of models can be transformed. What kind of Petri net model is the most ideal for functional partition shall be taken into consideration.

* Supported by the National Natural Science Foundation of China (61163012,61363074) ;the University Scientific Research Project of Guangxi(2013YB147)

Since Petri fundamental net is too simple to effectively simulate the behavior of the system, and the change of the indicator of the state place when the changes occur during the running of the actual system; While the other complex network systems like high-level Petri net and time Petri net which are expanded on the basis of P/T net are not the ideal model for functional partitioning. P/T model is the capacity function and the weight function with directional edge set by increasing place on the basis of the prototype Petri net, it can vividly describe the system's structure, behavior and the changes of state place indicators, and it is easier for other net models to transform into it. So, the P/T net is the ideal choice as the research object of functional partition. The transforming steps of Petri net structure model are as follows [1-3]:

There is no need to transform if net ***N*** is P/T net; Transform net ***N*** into P/T net model with equivalent structure to ***N*** if ***N*** is not P/T net. The transforming steps are as follows:

(1) Each place of ***N*** is transformed into the place of P/T net and the initial signs of the place keep the same;

(2) Each change of ***N*** is transformed into the changes of P/T net;

(3) Each arc of ***N*** is transformed into the arcs of P/T net and the weights on the arcs are the same.

III. PLACE INVARIANT AND SUBNET PARTITIONING METHOD PETRI NET PARALLEL SUBNETS DIVISION CONDITION

A. Solving of place invariant

Definition 1[4] Suppose $N=(P,T;F)$ is a network, $|P|=m$, $|T|=n$, D is the incidence matrix of ***N***. If there is a non-trivial m-dimensional nonnegative integer vector X which meets $DX=0$, then we say that X *is* a place invariant of net ***N***.

Definition 2[4] Suppose X *is a* place invariant of net $N=(P,T;F)$, then we say that $\|X\|=\{p_i \in P | X(i)>0\}$ is the support of place invariant X.

P/T net model is very complex and contains large numbers of place and changes, so it is very difficult to conduct manual calculation and verification. The following are the methods and steps for solving the place invariant with computers:

(1) Input of initial data. Initial data includes the output matrix $D^+=[d_{ij}^+]$ and input matrix $D^-=[d_{ij}^-]$of P/T net model, and initial sign $M=(M(p_1),M(p_2),\ldots,M(p_m))$. After the output matrix $D^+=[d_{ij}^+]$ and input matrix $D^-=[d_{ij}^-]$ are determined,

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.21

the incidence matrix $D=D^+ - D^-$ of Petri net model can be generated automatically by the computer.

(2) Obtain the rank r of the incidence matrix D. Suppose that $|P|=m$, $|T|=n$, then the incidence matrix D is a matrix with n rows and m columns. n and m represent the number of changes and the number of place of the P/T net respectively; X is a m-dimensional vector, which is the place invariant to be solved and represents the state of the place in the net. When $r=m$, the homogeneous linear equation set $DX=0$ only has zero solution, the Petri net doesn't have concurrent processes and can't conduct functional partition, the end; When $r<m$, turn to the next step;

(3) Solve the homogeneous linear equation set $DX=0$ and obtain the place invariant of P/T net. Use iterative method of elementary matrix transformation for solving the fundamental solutions to homogeneous linear equation set or the orthogonalizd column processing method to obtain the non-zero solutions and these non-zero solutions constitute the place invariant set Γ of P/T net.

(4) According to definition 2, obtain the place invariant supports from the place invariant set to generate with the place invariant support set $\Gamma_{p.}$

B. Subnet partitioning conditions of place invariant

Net $\boldsymbol{N}$ may have multiple place invariant X, then the place invariant set Γ becomes a possible partition of P/T net model. However, not every place invariant (subnet) can be the parallel subnet with independent function of the P/T net. It is also needed to examine the elements of the place invariant set, if they meet the following conditions of the theorem, then they can become the independent subnets of P/T net.

Theorem 1 Suppose that Γ_p is the place invariant support set of p/T net $\boldsymbol{N}= (P,T;F,K,W,M)$, the subnets corresponded to the elements of Γ_p meet the following conditions, then subnets corresponded to these elements are all a possible partition corresponded to net $\boldsymbol{N}$, as a result of partitioning, we can obtain a group of Petri net subnets with independent function[5].

$$\forall p\in\{p_i\in P|X(i)>0\},\forall p\in(t,t')\in p^{\bullet}\cap{}^{\bullet}p \tag{1}$$

$$W(p,t)=W(t',p)=1\wedge\sum_{p\in\|X\|}M(p)\geq 1 \tag{2}$$

$$\bigcup_{i=1}^{|\Gamma_p|}\Big(\bigcup_{j\in\|s_i\|}(p_j^{\bullet}\cup{}^{\bullet}p_j)\Big)=T \tag{3}$$

$$\forall s_1,s_2\in\Gamma_p\|s_1\|\cap\|s_2\|=\varnothing \tag{4}$$

In our study, we find that if the place signs of the subnets corresponded to the place invariant set are all zero, then the subnets don't meet the conditions of formula (2) in theorem 3; When the subnets of two place invariant sets have shared place, the two subnets don't meet the conditions of formula (4) either. For lacking of some subnets, although other subnets can create processes, the whole network system can't run normally. Therefore, we extend theorem 1 and get the completeness theorem 2 on subnet partitioning.

Lemma 1 Among the elements of P- invariant support set $\Gamma_{p.}$ if increase a entered non-null place for any change t of subnet N_s corresponded to all the elements $\|s\|$ of the place whose initial sign is null, then the formed extended subnet N_{s+p} can be partitioned into a process with independent function .

Prove: Increase a entered non-null place for any change t of subnet N_s, we can know through definition mark that a initial sign M must exist to make change t enabled, namely $M[t>$. The extended subnet N_{s+p} can run and can create the processes. The end.

Theorem 2 If the elements of the place invariant support set Γ_p of Petri net $\boldsymbol{N}= (P,T;F,K,W,M)$ meet the following conditions:

$$\forall p\in\{p_i\in P|X(i)>0\},\forall p\in(t,t')\in p^{\bullet}\cap{}^{\bullet}p \tag{5}$$

$$(W(p,t)=W(t',p)=1)\wedge\Big(\sum_{p\in\|X\|}M(p)\geq 1\vee\Big(\sum_{p\in\|X\|}M(p)=0\wedge(M({}^{\bullet}(p^{\bullet}))\geq 1\vee M({}^{\bullet}({}^{\bullet}p))\geq 1)\Big)\Big) \tag{6}$$

$$\bigcup_{i=1}^{|\Gamma_p|}\Big(\bigcup_{j\in\|s_i\|}(p_j^{\bullet}\cup{}^{\bullet}p_j)\Big)=T \tag{7}$$

$$\forall s_1,s_2\in\Gamma_p\|s_1\|\cap\|s_2\|=\varnothing \tag{8}$$

Then these elements can be partitioned into parallel subnets.

Prove: the theorem 1 and lemma 1 shows that the conclusion is obvious

C. Partitioning algorithm of subnets

According to the subnet partitioning process of P/T net given above, we obtain the subnet partitioning algorithm of place invariant-based P/T net. The steps are as follows:

(1) Transform the non- P/T-net Petri net model into P/T net model, if there are directional arcs from place to changes or from changes to place, then let the weights to be 1, and if there is no directional arc, then the weights are all 0;

(2) Input the initial data including output matrix $D^+=[d_{ij}^+]$, input matrix $D^-=[d_{ij}^-]$ of the Petri net model and initial sign $M=(M(p_1),M(p_2),\ldots,M(p_m))$;

(3) Solve the incidence matrix D through the formula $D=D^+ - D^-$;

(4) Solve the homogeneous linear equation set $DX=0$, and get the place invariant set of the P/T net;

(5) For each element of the place invariant set, get the set of place invariant supports which meets definition 2;

(6) Verify whether the subnets corresponded to all the elements in the place invariant support set meet the conditions of (6) -(8) in theorem 2. If they meet, then the creating process can be obtained; otherwise, it ends.

(7) The place invariant support which meets the conditions of (6)-(8) in theorem 2 is the parallel subnet of P/T net.

IV. BASED ON MPI + OPENMP + STM PETRI NET PARALLEL ALGORITHM OF THREE LAYER PARALLEL MODEL

A. The MPI + OpenMP + STM three layer parallel model

In order to make full use of multi-core clusters architecture, we put every piece of multi-core PC node as a process, cluster platform of multiple nodes, constituting a multiple processes. The data transfer in process realizes through the MPI function place; MPI parallel indicates nodes parallel; multi-threading is founded by OpenMP in each node of clusters and mapped to multi-core processors, concurrent sharing memory between threads is achieved by transactional memory (STM), node in parallel is mainly for STM concurrent execution; thus formed MPI + OpenMP + STM layer 3 parallel programming model among multiple processes and within internal processes multiple threads that concurrent conducting, as shown in figure 1[6]. As each node within different threads access the same storage unit or critical region, use transaction memory instead of locking mechanisms, and conduct concurrently by multiple threads. Soft transactional memory (STM) model of the system is shown in figure 2[6].

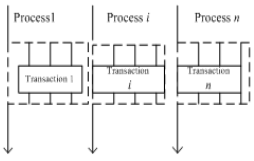


Fig. 1 Multi core cluster architecture

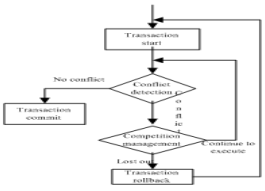


Fig. 2 STM model

B. The in-process multithreaded change behavior analysis

After P/T net division, each subnet is equivalent to a process, process includes the place as a multithreaded execution status or the state of resources, change is multithreaded process to perform an action or operation behavior. Subnet internal behavior analysis is as follows:

The first kind of situation: if the process change t, meet $|\cdot t|=|t\cdot|=1$, the change is only one input and one output place, this change is the local behavior of the process or local action;

The second kind of situation: if any two changes in the process of the t_1 and t_2, there is a logo M, making $M[t_1\rangle M_1 \rightarrow M_1[t_2\rangle$ and $M[t_2\rangle M_2 \rightarrow M_2[t_1\rangle$, the changes of the in-process t_1 and t_2 are concurrent, namely the occurrence of two changes will not influence each other;

The third kind of situations: if any two changes in the process of the t_1 and t_2, there is a logo M, making $M[t_1\rangle M_1 \rightarrow \neg M_1[t_2\rangle$ and $M[t_2\rangle M_2 \rightarrow \neg M_2[t_1\rangle$, t_1 and t_2 are in-process changes in ***M*** conflict. In the conflict of the two changes, when one occurs, another change will lose happened right, and vice versa.

C. In-Process change behavior analysis

If the process is shared between changes, namely $\Gamma_p \cap \Gamma_d = \{t_i\} \neq \varnothing$, it indicates that two processes have more than one input and output between the changes of the place, there is synchronous change at these two processes. The change playing the role of a process server, through the process server, the exchange of information between the two processes can be conducted, namely the MPI function is applied to implement the exchange of information. Specific implementation, the two processes is, in fact, conducting the exchange of information with a process server[7-9].

D. Parallel algorithm of Petri net of MPI + OpenMP + STM three- layer parallel model

According to the analysis above, the parallel algorithm of Petri net based on multi-core clusters is as follows:

Input: Input matrix and output matrix of Petri net;

Output: Parallel subnets of P/T net.

Begin:

Step1 Enter the input matrix and output matrix of Petri net from the master process;

Step2 Call and execute the parallel function of Petri net model transforming into P/T net model;

Step3 Call and execute the parallel function of solving the place invariant and support of P/T net model;

Step4 Call and execute the parallel function of parallel subnet partitioning algorithm of P/T net;

Step5 Find the shared changes among the subnets according to the condition $\Gamma_p \cap \Gamma_d = \{t_i\} \neq \varnothing$, and take each shared change as a process server;

Step 6 Have several processes mapped to different machines on the platform of multi-core clusters respectively, and use MPI to send and receive the function data. After each machine receives the data from the master process, then divide the data into several parts, each part is processed by a core (thread) and the result is sent back to the master process from the slave process. If conflict occurs to the data, then use STM transactional memory to process;

Step7 The function realization of P/T process. Each process is a sequence state machine which can be viewed as a programming task model, when it reaches the non-subsequent state place, the loop ends;

Step8 The function realization of the internal change concurrency and conflict of the process. No matter the internal change is executed concurrently or sequentially, these changes will be viewed as the local action of sequential running, adopting multi-thread parallel programming within the same process and parallel execution. If conflict occurs to several changes, then add place for them, these changes turn to parallel execution from conflict after forming a control loop, and the processing process comes down to the concurrent processing of changes with STM transactional memory; Go to step 3 for execution;

Step9 The function realization of the shared change buffer between two processes. The functions of the change include the local actions, such as receiving information from two connection processes, information processing and sending information to the two processes connected with it.

Step10 Output incidence matrix, place invariant support and parallel subnets in the master process.

The End.

V. APPLICATION EXAMPLE AND EXPERIMENT RESULT ANALYSIS

Finally, examine the parallel algorithm of Petri net system through a simulated deposit and withdrawal system of the bank (see figure 3). The relevant specific parameters of this system are as follows:

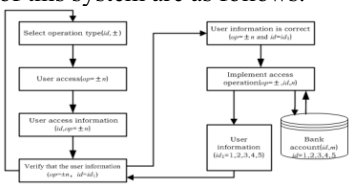


Fig. 3 Bank access data stream

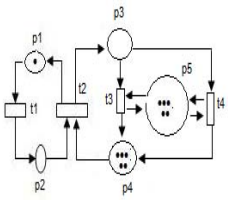


Fig. 4 Bank access data stream Petri net

- id represnets user account, m represents the amount of the user's deposits, m can be different for different users;
- The database stores user information with (id, m), which indicates the current amount of deposit in the bank of the user id;
- Operation-type symbol "+" represents depositing, "-" represents withdrawing, "±" represents "+" or "-";
- op represents operation, *op*=+ or *op*=- respectively represents the operation of depositing or withdrawing;
- (op，id，n) represents that the user whose account is id conducts the operation of depositing n Yuan or withdrawing n Yuan;
- *op*=± n and id=id1 indicate that the user account and bank account are the same, and it is available for the user whose account is id to conduct the operation of depositing or withdrawing;
- Suppose that there are 5 users, id = 1, 2, 3, 4, 5.

In order to solve the actual problems above, refer to the data flow of deposits and withdrawals in figure 3 and the Petri net model of the parameter structure system as shown in figure 4. Wherein, P1 represents the type of operation, p2 represents the information of user depositing and withdrawing, p3 represents that user information is correct, p4 represents user id information and p5 represents bank account place; T1 represents user operation of depositing and withdrawing; t2 represents receiving user id and verifying user information, t3 represents the machine's operation of automatic depositing and t4 represents the machine's operation of automatic withdrawing.

On the platform formed by 8 PCs of Linux9.0 operating system, 4-core CPU and 1 GB of memory capacity, the experimental process and results are as follows:

Input the initial data of Petri net model including: the output matrix D^+ [4][5] = {{0,1,0,0,0}，{1,0,1,0,0}，{0,0,0,1,1}，{0,0,0,1,1}}, input matrix D^-[4][5] = {{1,0,0,0,0}，{0,1,0,1,0}，{0,0,1,0,1}，{0,0,1,0,1}}, and initial sign vector M_0=(1,0,0,5,5). Conduct depositing and withdrawing to 100, 1,000 and 10,000 user processes respectively randomly, and the program running result is correct.

Experimental result and analysis: The place invariants are respectively X^T_1= (1, 1, 0, 0, 0)，X^T_2=(0, 0, 1, 1, 0)，X^T_3=(0, 0, 0, 0, 1); X_1, X_2 and X_3 are supports; the subsets Γ_{p1}={X_1、X_3} and Γ_{p2}={X_1、X_2} meet the conditions of theorem 4. So, the simulated bank depositing Petri net system has two kinds of functional partition, one kind is partitioning it into Pa process and Pb process according to the subnets corresponded to X_1 and X_2. The other kind is partitioning it into Pc process and Pd process according to the subnets corresponded to X_1 and X_3. The simulated experimental result shows that the Petri net system can conduct subnet partitioning with the method of place invariant, and be designed according to the functional partition result process parallel algorithm, to make the Petri net system conduct parallel running and realize Petri net modeling.

VI. CONCLUSIONS

This paper conducts subnet partitioning of Petri net system by using place invariant technology, puts forward the parallel algorithm of Petri net based on multi-core clusters combining with the MPI+OpenMP+STM three layer parallel model, and simulates the parallel running process of Petri net through the examples. The simulation experiment shows that the parallel algorithm of Petri net system place invariant is feasible and effective, and the concrete application of this algorithm in related fields will be the focus of future research.

REFERENCES

[1] M. Paludetto. Sur la commande de procedes industriels:unemethodologie basee objets et reseaux de Petri[D]. These de doctorat,Universite Paul Sabatier,Toulouse, France,1991,pp.34-47.

[2] W.El Kaim and F.Kordon. An integrated framework for rapid system prototyping and automatic code distribution[C]. In 5thIEEE International Workshop on Rapid System Prototyping,Grenoble,IEEE Comp.Soc.Press,1994,pp.52-61.

[3] J.M.Colom and M.Silva. Convex geometry and semiflows in P/T nets. A comparative study of algorithms for computation of minimal P-semiflows[J]. In Rozenberg,Advances in Petri Nets 1990,Volume 483 of Lecture Notes in Computer Science.Springer-Verlag,1991,pp.79-112.

[4] WU Z H.Petri Nets Introduction[M]. Beijing: Mechanical industry publishing house.2006, pp.144-155.

[5] Girault C and Valk R.Petri Nets for Systems Engineering: A Guide to Modeling, Verification, and Applications [M].Springer-Verlag Berlin Heidelberg. 2003,pp.159-235.

[6] LI Wen-jing,LI Shuang1,YUAN Chang-an.Research on Software Transactional Memory Parallel Programming Model Based on Multicore Cluster[J]. Journal of Chinese Computer Systems,2014,35(8),pp.1732-1734.

[7] Hao Kegang, Ding Jianjie. Hierarchical Petri nets[J].Journal of Frontiers of Computer Science and Technology,2008, 2(2): 123-130.

[8] Xuling Chang,Linpeng Huang,Jianpeng Hu.Real-time reconfiguration of distributed control system based on hard Petri nets[C]. 2014 IEEE 38th Annual Computer Software and Applications Conference (COMPSAC) 2014, pp. 267-272.

[9] Martinez-Araiza, U. ,Lopez-Mellado, E. A CTL model repair method for Petri Nets[C].World Automation Congress (WAC), 2014 ,pp: 654 – 659.

Study and Realization on the Partitioning Algorithm of Parallel Subnet of Petri Net System

Wen-jing LI
School of Logistics Management and Engineering
Guangxi Teachers Education University
Nanning, China
e-mail:liwj@gxtc.edu.cn

Xiang-bo Zhang, Yingzhou BI , Xuan Wang
College of Computer and Information Engineering
Guangxi Teachers Education University
Nanning, China

***Abstract*—In order to solve the problem about the partition of Petri net model and subnet division, realize the concurrent execution or simulation runs of Petri net system, the partitioning algorithm of parallel subnet of Petri net is proposed. First, as Petri net system has the characteristics of synchronization and concurrence, provide the** place**-invariant-based Petri net model partitioning and subnet division conditions and parallelizing analysis; put forward the extended theorem and validation of partitioning condition of parallel subnet; then, provide the formalization of subnet division and the solving process of place invariant and** place**-invariant-based partitioning algorithm of parallel subnet of Petri net. The experimental results show that the partitioning algorithm of parallel subnet of place-invariant-based Petri net is feasible and effective.**

Keywords- Petri net; place invariant; parallel subnet; partitioning conditions; partitioning algorithm

I. INTRODUCTION

At present, people mainly conduct static analysis and research on the structure, behavior and function of all kinds of system models, such as prototype Petri net, colored Petri net, time Petri net and so on. But the dynamic properties of the actual system like behaviors and functions need to be reflected through the simulation, animation or running of the system. So, conducting research on parallel execution or simulation runs of actual Petri net system has very important significance. During the study on the parallelization of Petri net system, concentration method is mentioned in the P/T net in literature [1]. This method scan each change in the model to check the trigger of the transition, but it can't keep the parallelism of the model; decentralized method is mentioned in colored Petri net in literature [2]. This method completely focuses on distributed execution, and every place and change are implemented through the process. It keeps the parallelism of the model, but its efficiency will become very low when the network scale gets large or number of the color collection element is high. Parallelization method of place invariant is mentioned in literature [3], but this method is only applicable to the situation when there are positive place in the subnet, it also has two shortages: first, it doesn't give the parallelizing conditions of different subnets not given subnet when the subnets are all empty place; second, the parallelism problem of the Petri net when there is no place invariant. On the basis of literature [3], expand the partitioning conditions of subnet of Petri net, to realize the partitioning algorithm of parallel subnet of place-invariant-based Petri net.

* Supported by the National Natural Science Foundation of China (61163012) ;Guangxi Natural Science Foundation (2012GXNSFAA053218); the University Scientific Research Project of Guangxi(2013YB147)

II. PARTITION OF PARALLEL SUBNETS OF PETRI NET SYSTEM

Petri net is a model used to describe distributed system. It can describe the structure of the system as well as simulate the running of the system. See Petri net and concepts and terminology related to this paper in literature [4-5].

A. Selection of net model

It is needed to select a suitable network structure model first to realize the parallelization of Petri net model to partition the network into several subnets[4][5], then form all the subnets as a group of processes, and these processes are created as parallel process respectively.

The reason for choose P/T net structure model. There are differences among Prototype network, P/T net and high-level network as well as close inner links. In network structure models, the prototype net model can become simpler P/T net model through abstract folding, and the P/T net model can become simpler high-level net model through further abstract folding. Whereas, the relatively simple high-level network system model can also be transformed into corresponding P/T net model. In algebraic model, the prototype network is the basic network, P/T net is established by increasing capacity function and weight function on the basis of prototype network; and the high-level network is classifying to-ken on the basis of P/T net, capacity function and weight function turn to K-dimensional nonnegative mapping function of high-level network from one-dimensional nonnegative mapping function of P/T net. A complex system can be described using the prototype Petri net, P/T net or high-level network system, though the complexity of the models are different, they have similarity in simulating the performance of the actual system, which means that the simulating capability of the high-level network is not stronger than that of P/T network and prototype network, and the simulating capability of P/T net

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.22

is not stronger than that of prototype network[5]. As P/T net is closer to the running of actual system, P/T net structure is the ideal structure model of partitioning subnets.

B. Transformation method of network structure model

Transformation method of network structure model. There is no need to transform if net N is P/T net; transform net N into P/T net model with equivalent structure to N if N is not P/T net. During the transformation, (1) each place of N shall be transformed into the place of P/T net and the initial signs of the place keep the same; (2) Each change of N shall be transformed into the transition of P/T net; (3) each arc of N shall be transformed into the arcs of P/T net and the weights on the arcs are the same.

C. Partitioning conditions of place invariant-based subnets

Partition the P/T net applying place invariant, and stipulate the partitioning conditions of parallel subnet and subnets from the place invariant. The place invariant is defined as follows:

Definition 1[4] Suppose N= (P, T; F) is a network, $|P|=m$, $|T|=n$, D is the incidence matrix of N. If there is a non-trivial m-dimensional nonnegative integer vector X which meets $DX=0$, then we say that X *is* a place invariant of net N.

Definition 2[4] Suppose X *is a* place invariant of net N= (P, T; F), then we say that$\|X\|=\{p_i \in P|X(i)>0\}$is the support of place invariant X.

It can be learned from definition 1-2, that net N may have multiple place invariant X, and that all the place invariant of net N form a set Γ. Among them, the place invariant set Γ represents a possible partition in the net model. The place corresponded to the nonzero elements of each place invariant vector X and the extended set of the place constitute the partitioned subnet, and the subset is $N_i=\bigcup_{\{p_i \in P|X(i)>0\}} p_i \cup {}^\bullet p_i \cup p_i^\bullet, i=1,2,\cdots,m$ .Although net N is partitioned into several subnets according to the place invariant, however, not every place invariant (subnet) can become a parallel process. It is needed to investigate the set Γ_p composed of the supports of place invariant Γ to decide which place invariant needs to set parallel process, and the subnet corresponded to the element is $N_i=\bigcup_{\{p_i \in P|X(i)>0\}} p_i \cup {}^\bullet p_i \cup p_i^\bullet, i=1,2,\cdots,m$. If the elements of Γ_p meet the conditions of (1-4) in theorem 1, then the subnets corresponded to the elements can be set to be the parallel processes of net N .

Theorem 1 In the place invariant support set Γ_p of Petri net N=(P, T; F, K, W, M) , the subnets corresponded to the elements meet the following conditions:

$$\forall p \in \{p_i \in P|X(i)>0\}, \forall p \in (t,t') \in p^\bullet \cap {}^\bullet p \tag{1}$$

$$W(p,t) = W(t', p) = 1 \wedge \sum_{p \in \|X\|} M(p) \geq 1 \tag{2}$$

$$\forall s_1, s_2 \in \Gamma_p \|s_1\| \cap \|s_2\| = \varnothing \tag{3}$$

$$\bigcup_{i=1}^{|\Gamma_p|} \left(\bigcup_{j \in \|s_i\|} (p_j^\bullet \cup {}^\bullet p_j) \right) = T \tag{4}$$

Then the subnets are parallel subnets of net N.

Prove:

(1) If the place invariant support of net N only has one element and it meets the condition of (1)-(4), then the corresponding subnet to the element is a parallel subnet;

(2) Suppose that the place invariant support set Γ_p of net N has n elements. It is known that the subnets corresponded to n-1 elements are parallel subnets of net N, and the n-1 elements meet the conditions of (1)-(3) while don't meet the condition of (4). The nth element also meets the conditions of (1)-(3), it can be known according to the condition of (4) that the n-1 elements and the nth element constitute the place invariant support set Γ_p of net N, and they meet the condition of (4). The end.

D. Extension of partitioning conditions of subnets

Some subnets corresponded to the place invariant supports don't meet the conditions of (2) in theorem 1 and can not be partitioned into parallel subnets. For lacking of subnets, P/T net cannot meet the condition of (4) in theorem 1 overall. The absence of subnet makes the whole network system not run normally. So, we extend the conditions[6] of (2) in theorem 1 and get theorem 2.

Lemma 1 In the elements of the place invariant support set Γ_p of Petri net N=(P,T;F,K,W,M), if increase a entered non-null place for any change t of subnet N_s corresponded to all the elements $\|s\|$ of the place whose initial sign is null, then the formed extended subnet N_{s+p} can become a parallel subnet of net N.

Prove: Increase a entered non-null place for any change t of subnet N_s, a initial sign M must exist to make change t enabled, namely $M[t>$. The extended subnet N_{s+p} can run and can become a parallel subnet of net N. The end.

Theorem 2 If the elements of the place invariant support set Γ_p of Petri net N=(P,T;F,K,W,M) meet the following conditions:

$$\forall p \in \{p_i \in P|X(i)>0\}, \forall p \in (t,t') \in p^\bullet \cap {}^\bullet p \tag{5}$$

$$(W(p,t) = W(t', p) = 1) \wedge \left(\sum_{p \in \|X\|} M(p) \geq 1 \vee \left(\sum_{p \in \|X\|} M(p) = 0 \wedge (M({}^\bullet(p^\bullet)) \geq 1 \vee M({}^\bullet({}^\bullet p)) \geq 1) \right) \right) \tag{6}$$

$$\forall s_1, s_2 \in \Gamma_p \|s_1\| \cap \|s_2\| = \varnothing \tag{7}$$

$$\bigcup_{i=1}^{|\Gamma_p|} \left(\bigcup_{j \in \|s_i\|} (p_j^\bullet \cup {}^\bullet p_j) \right) = T \tag{8}$$

Then the subnets or extended subnets corresponded to the elements are all the parallel subnets of net N.

Prove: It can be known from theorem 1 and lemma 1 that the conclusion is obvious.

For example, the incidence matrix of Petri net model is as follows:

$$\begin{bmatrix} -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Then $X_1=[0, 1, 1, 0, 0, 0, 0, 0]^T$ 、 $X_2=[0, 0, 0, 1, 1, 0, 0, 0]^T$、 $X_3=[0, 0, 0, 0, 0, 0, 1, 1]^T$ is the place invariant of net N; and $||X_1||=\{p_2, p_3\}$ 、 $||X_2||=\{p_4, p_5\}$ 、 $||X_3||=\{p_7, p_8\}$ is also the place invariant support of net N. Partition net N into three subnets in accordance with the part corresponded to the place invariant X_1, X_2 , X_3, the three subnets obviously meet the conditions of (5)-(8) in theorem 2 and are the parallel subnets of net N. So, their corresponding subnets or expanding subnets can create parallel processes.

III. THE PARTITION ALGORITHM OF PARALLEL SUBNETS

In general, after establishing Petri net model for actual problems, it is needed to formalize the behavioral norms of the net system and to solve and verify the place is invariant to realize the parallel running of the network[7].

A. Formalized process

The formalized process of P/T net is as follows:

(1) Regard all the place and transition of the net model as two different kinds of vertices of the graph. The vertex set of the place is represented with $P=(p_1,p_2,...,p_m)$, the vertex set of the transition is represented with $T=(t_1,t_2,...,t_n)$;

(2) If there is directional arc from the change to the place and the weight is $W(t,p)$, then it is represented with $W(p,t)$. Conversely, it will be represented with 0. But in parallel subnet partitioning, only consider the situation when the weight is 1 (the same below), let $W(t,p)=1$. We can get the output matrix of the net model as $D^+=[d_{ij}^+]$;

(3) If there is vector arc from the change to the place and the weight is $W(t,p)$, then it is represented with. $W(p,t)=1$. Conversely, it will be represented with 0. We can get the input matrix of the net model as $D^-=[d_{ij}^-]$;

(4) The number of the initial signs of the net place is represented with initial signs. If the net model contains m place, namely $|P|=m$, then the initial sign of m-dimensional vector M is $(M(p_1),M(p_2),\ldots,M(p_m))$.

B. Solving of place invariant

If the P/T net model is very complex and contains large numbers of place and transition, so it is very difficult to conduct manual calculation and verification. Therefore, provide the following procedures for automatically solving and verifying the place invariant:

(1) Input of initial data. Initial data includes the output matrix $D^+=[d_{ij}^+]$ and input matrix $D^-=[d_{ij}^-]$of P/T net model, and initial sign $M=(M(p_1),M(p_2),\ldots,M(p_m))$. After the output matrix $D^+=[d_{ij}^+]$ and input matrix $D^-=[d_{ij}^-]$ of are determined, its incidence matrix $D=D^+ - D^-$ can be generated automatically.

(2) Solve homogeneous linear equation set $DX=0$, and calculate the place invariant of P/T net. According to the structural features of the P/T net model, suppose that $|P|=m$, $|T|=n$, then the incidence matrix is a D with n rows and m columns. n and m represent the number of transition and the number of place of the P/T net respectively; X is a m-dimensional vector, which is the place invariant to be solved and used to show the state of the place in the net. And the homogeneous linear equation set of P/T net place invariant has three situations:

The first kind of situation: When $m>n$, which means that the number of Petri net place is more than the number of transition, then the number of unknowns for the homogeneous linear equation set is more than the number of equations;

The second kind of situation: When $m=n$, which means that the number of place equals to the number of transition, then the number of unknowns for the homogeneous linear equation set equals to the number of equations;

The third kind of situations: When $m<n$, which means that the number of Petri net place is less than the number of transition, then the number of unknowns for the homogeneous linear equation set is less than the number of equations;

(3) Judge whether the elements of the place invariant support set meet the partitioning conditions of parallel subnet in theorem 2.

(a) Judge whether the sets meet the two conditions of formula (6) one by one, it can be considered that the subnets corresponded to the set may constitute parallel subnets as long as the set meets one of the conditions, otherwise, judge the nest set.

(b) Among all the sets which meet the conditions of formula (6), conduct pairwise comparison according to the conditions of formula. It is easy to conduct that as it only needs to judge that every two sets have no common elements, namely shared place.

(c) For all the sets which meet the conditions of formula (6) and formula (7), the precursor and subsequent transition (only select one if there are same transition) of their non-zero place constitute set T_{p+}, and if $T_{p+}=T$, then the subnets corresponded to these supports can create parallel processes and output the place invariant set vector; and if$T_{p+}\neq T$, then output the information about the place invariant which does not meet the conditions.

C. Partitioning algorithm of place invariant-based parallel subnets

According to the theorems about partitioning conditions of parallel subnet of P/T net and the process of formalization and solving the place invariant given above, we get the partitioning algorithm of parallel subnets of place invariant-based P/T net. The steps are as follows:

Input: Input matrix $D^+=[d_{ij}^+]$, $D^-=[d_{ij}^-]$ and initial sign $M=(M(p_1),M(p_2),\ldots,M(p_m))$;

Output: Parallel subnets.

Begin:

Step1: Transform the non- P/T-net Petri net model into P/T net model, if there are directional arcs from place to transition or from transition to place, then let the weights to be 1, and if there is no directional arc, then the weights are all 0;

Step2: Input the initial data including output matrix $D^+=[d_{ij}^+]$, input matrix $D^-=[d_{ij}^-]$ of the Petri net model and initial sign $M=(M(p_1),M(p_2),\ldots,M(p_m))$;

Step3: Solve the incidence matrix D through the formula $D=D^+ - D^-$;

Step4: Solve the homogeneous linear equation set $DX=0$, and get the place invariant set of the P/T net;

Step5: For each element of the place invariant set, get the set of place invariant supports which meets definition 2;

Step6: Partition P/T net according to the elements in the place invariant support set and get corresponding subnet;

Step7: Verify whether the subnets corresponded to all the elements in the place invariant support set meet the conditions of (6) and conditions of (7) in theorem 2. If they meet, then the parallel subnets can be obtained.

The End

IV. EXPERIMENT AND RESULT ANALYSIS

We write the C program of parallel subnet partitioning of place invariant-based P/T net, the program consists of three functions including data input function, place invariant function and verification function. Among them, the data input function is responsible for the input of initial data, such as output matrix $D^+=[d_{ij}^+]$, input matrix $D^-=[d_{ij}^-]$ and initial sign $M=(M(p_1),M(p_2),\ldots,M(p_m))$;place invariant function is responsible for solving the homogeneous linear equation set, place invariant and the set; verification function is responsible for verifying whether the place invariant support set meet the conditions of formula (6), formula (7) and formula (8) in theorem 2.

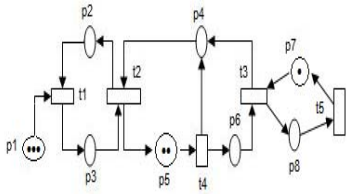


Fig.1 P/T net

Take P/T net model in figure 1 as an example, conduct experiment verification in VC++6.0 environment. The entered out matrix is: $D^+[5][8]$ = {{0,0,1,0,0,0,0,0} , {0,1,0,0,1,0,0,0} , {0,0,0,1,0,0,0,1} , {0,0,0,1,0,1,0,0},{0,0,0,0,0,0,1,0}}; the input matrix is D^- [5][8] = {{1,1,0,0,0,0,0,0} , {0,0,1,1,0,0,0,0} , {0,0,0,0,0,1,1,0}, {0,0,0,0,1,0,0,0},{0,0,0,0,0,0,0,1}}; and the initial sign vector isM_0=(3,0,0,0,2,0,1,0).

Experiment result: the place invariants are X^T_1= (0,1,1,0,0,0,0,0) , X^T_2=(0,0,0,1,1,0,0,0) , X^T_3=(0,0,0,0,0,0,1,1); X_1、X_2、X_3 are supports; X_1、X_2、X_3 meet the conditions of theorem 2 and are all parallel subnets.

The experiment result is consistent with the theoretical analysis in section 1.4. So, the partitioning algorithm of parallel subnets of place invariant-based Petri net system we put forward is feasible and effective, and it is suitable for the parallelization of complex Petri net system, such as specific distributed parallel processing, discrete events and flexible manufacturing. Acquire the parallel subnets of Petri net system according to this approach, then make each subnet map to different processors of the parallel platform, and program for the behavior, function and operation of each process (subnet)[8] , and then the Petri net system can run.

V. CONCLUSIONS

This paper provides the partitioning algorithm of parallel subnets of place invariant-based Petri net, according to which determine the number of parallel process Petri net creates, determine the transition, place and their behavior each parallel subnet contains, and program for their realization. This is an effective method of realizing the partition of Parallel subnet of Petri net, but it also has its limitations, for example, when the rank $r(D)=m$ in the homogeneous linear equation set $DX=0$, there is only one zero solution and the Petri net doesn't have place invariant. In this case, how to conduct the partitioning and parallelization of the parallel subnet of the Petri net is our main research work next.

REFERENCES

[1] M. Paludetto. Sur la commande de procedes industriels:unemethodologie basee objets et reseaux de Petri. These de doctorat,Universite Paul Sabatier,Toulouse, France,1991 : 34-47.

[2] W.El Kaim and F.Kordon. An integrated framework for rapid system prototyping and automatic code distribution. In 5thIEEE International Workshop on Rapid System Prototyping,Grenoble,IEEE Comp.Soc.Press,1994:52-61.

[3] J.M.Colom and M.Silva. Convex geometry and semiflows in P/T nets. A comparative study of algorithms for computation of minimal P-semiflows. In Rozenberg,Advances in Petri Nets 1990,Volume 483 of Lecture Notes in Computer Science.Springer-Verlag,1991:79-112.

[4] WU Z H.Petri Nets Introduction[M]. Beijing: Mechanical industry publishing house.2006, pp.144-155.

[5] Girault C and Valk R.Petri Nets for Systems Engineering: A Guide to Modeling, Verification, and Applications [M].Springer-Verlag Berlin Heidelberg. 2003,pp.159-235.

[6] Wenjing LI, Shuang LI,Shuju Li. Study on Function Partition Strategy of Petri Nets Parallelization. 12th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Publicshed by IEEE conference publishing services ,2013,pp.89-94.

[7] Fei Liu,Monika Heiner.Petri Nets for Modeling and Analyzing Biochemical Reaction Networks, Approaches in Integrative Bioinformatics ,2014, pp.245-272.

[8] Bartosz Jasiul, Marcin Szpyrka, Formal Specification of Malware Models in the Form of Colored Petri Nets.Lecture Notes in Electrical Engineering Volume 330, 2015, pp. 475-482 .

Temporal Logic of Stochastic Actions for Verification of Probabilistic Systems

LI Jun-tao
Information College
Guizhou University of Finance and Economic
Guiyang, china
E-mail: jtxq@qq.com

LONG Shi-gong,
Computer Science and Technology College
Guizhou University
Guiyang, china
E-mail: 526796467@qq.com

Abstract-The specification and verification of probabilistic systems were usually based on Computational Tree Logic, and systems and properties were specified by different language respectively. This paper extends and reforms Temporal Logic of Actions, puts foreword Temporal Logic of Stochastic Actions (TLSA), which can use additional state-action probabilistic distribution and probabilistic operator to specify probabilistic systems and their properties in the same logic.

Keywords-Specifying systems; System verification; TLA; Probabilistic systems

I. INTRODUCTION

As soon as the Model Checking was invented in 1980's, researchers had started applying it to the study of Probabilistic Systems' verification. In the beginning, people focus on the qualitative properties of system, for example, the program is whether or not terminating with probability 1. Afterwards, the algorithms for verifying quantitative properties have been progressed also. At present, the verification technology of Probabilistic Systems is mainly used for the field of security, distributed algorithms, systems biology, and system performance analysis, and so on.

In past thirty years, the implementation model of Probabilistic Systems is mainly based on Markov decision processes [1][2](MDPs), there are also some others import timed automata[6], putdown automata[7], or two-player game. They specify the property of system with linear temporal logic (LTL), ω-regular properties or probabilistic computation tree logic (PCTL), the latter imports probability distribution to computation tree logic (CTL); it is most used description language of Probabilistic Systems. Obviously, the traditional research method used different description language to specify the models or properties of Probabilistic Systems; this is not well for the property verification and design implementation of system.

This paper puts forward Temporal Logic of Stochastic Actions (TLSA), it is extension of Temporal Logic of Actions [4] (TLA) with probability, the latter is based on linear temporal logic, it defined the actions and operators, and can achieve specifying and verification of concurrent systems and their properties; the former inherit the most prominent feature of TLA: system and its properties can be described using TLSA at the same time.

II. PROBABILISTIC TRANSITION SYSTEM

Probabilistic Transition Systems (PTSs) is abstract model of Probabilistic Systems, its definition is followed:

Defination1.1 Probabilistic Transition systems is a 5-tuple: $\mathcal{P} = \{\mathcal{S}, \mathcal{I}, \mathcal{A}, \delta, \lambda\}$,among them:

$\mathcal{S}$：States set of system;

$\mathcal{I}$：Initial states set of system;

$\mathcal{A}$：Actions set of system;

δ ： $\mathcal{S}\times\mathcal{A}\rightarrow\mathcal{S}$ relationship of states trasation;

λ： （$\mathcal{S}\times\mathcal{A}\rightarrow\mathcal{S}$）$\rightarrow[0,1]$ is probability distribution of system actions, and meet the condition: $\forall s \in \mathcal{S}$, $\sum_{A\in\mathcal{A}} p(s,A,s') = 1$.

We can see that probabilistic transition systems are Label Transition Systems (LTS) adding a probability distribution of actions.

III. SYNTAX AND SEMANTICS OFTLSA

3.1 Syntax

TLSA's symbol includes:

（1） Probabilistic values: $pr(\in[0,1])$;
（2） Constant symbol: c, c_1, c_2 ...
（3） Rigid variables: u, u_1, u_2 ...
（4） Flexible variable: x, x_1, x_2 ...
（5） Atomic proposition: p, p_1, p_2 ...
（6） Constant element symbol: m, m_1, m_2 ...
（7） Arithmetic operator: +,-, *;
（8） Relational operators: <, =;
（9） logical operator: $\wedge$, $\neg$, $><_{pr}$
（10） Quantifier: $\exists$;
（11） Temporal operator：′, [].

The other conjunctions, for example $\vee$,

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.23

$\Rightarrow$, $\Leftrightarrow$, and so on, can be defined by $\wedge$ and $\neg$; Temporal operator $\Diamond$ can be defined by $\Box$ and $\neg$.

Defination2.1 *State:* Once assignment for all variables in system. All those possible assignments constitute the states set of system, named St. Initial states of system Φ write as $Init_\Phi$.

Defination2.2 *State functions* [4]:

$$f \triangleq c \mid u \mid x \mid f_1 + f_2 \mid f_1 - f_2 \mid f_1 * f_2$$

c is a constant, u is a rigid variable, x is a flexible variable.

Defination2.3 *State predicates* [4]:

$$q \triangleq p \mid (f_1 = f_2) \mid (f_1 < f_2) \mid \neg q \mid (q_1 \wedge q_2) \mid \exists u q$$

P is a atomic proposition, f_1 and f_2 are states functions, u is a rigid variable.

Defination2.4 *Action functions*:

$$f \triangleq c \mid u \mid x \mid x' \mid (f_1 + f_2) \mid (f_1 - f_2) \mid (f_1 * f_2)$$

c is a constant, u is a rigid variable, x is a flexible variable, x' is the value of x in new state, , f_1 and f_2 are functions of actions.

Defination2.5 *Actions*[4]:

$$A \triangleq p \mid p' \mid (f_1 = f_2) \mid (f_1 < f_2) \mid \neg A \mid (A_1 \wedge A_2) \mid \exists u A$$

p is a atomic proposition, p' is the value of p in new state，f_1 and f_2 are functions of actions, u is a rigid variable.

PS: All actions constitute the action set of system，named 。

Defination2.6 *State actions probability distribution:* if

$$p(s, A) \triangleq \begin{cases} pr, A \in \mathcal{A} \text{ 且 } s \in St \\ 0, \text{ 其它} \end{cases}$$

for $\forall s \in \mathcal{S}$ meet

$$\sum_{A \in \mathcal{A}} p(s, A) = 1$$

then $p(s, A)$ is state actions probability distribution. We write state actions probability distribution of present state as $p(\cdot)$, and the probability of action A at present state as $p(\cdot, A)$.

Defination2.7 *Probabilityoperator* $\succeq_{pr}$: is a probability range of a predicate, action, or another Boolean expression, $\succeq$ is one of $<$, $=$, $\geq$, $\leq$, $>$, $pr \in [0,1]$ is probability value. For example, $_{=0.4}(A)$ represents action A occur by probability of 0.4.

Defination2.8 *Probabilistic Actions*:

$$M \overset{p(\cdot)}{=} A_1 \vee A_2 \vee \cdots \vee A_n$$

$p(\cdot)$ is actions probability distribution of present state, A_i（$i \in$ ）is action,

Defination2.9 *Stuttering*:

$$[A]_v \triangleq A \vee (v' = v) \qquad (2.1)$$

$$\langle A \rangle_v \triangleq A \wedge (v' \neq v) \qquad (2.2)$$

A is a action, v is a tuple constituted by state variable, v' is value of v in next state.

$$[M]_v \triangleq M \vee (v' = v) \qquad (2.3)$$

$$\langle M \rangle_v \triangleq M \wedge (v' \neq v) \qquad (2.4)$$

Defination2.10 *Simple TSLA formula*:

$$F \triangleq P \mid \Box P \mid \Box [A]_v \mid \neg F \mid (F_1 \wedge F_2)$$

P is state predicate, A is a action.

Defination2.11 *Enabled*:

If an action A, its probability satisfy $p(\cdot, A) > 0$, then A is enable at present state., marked *Enabled*$\langle A \rangle_v$. *Enabled* is called Enable Predicate.

Defination2.12 *Fairness*:

$$WF_v(A) \triangleq \Box\Diamond \neg ENABLED \langle A \rangle_v \vee \Box\Diamond \langle A \rangle_v$$

$$SF_v(A) \triangleq \Diamond\Box ENABLED \langle A \rangle_v \vee \Box\Diamond \langle A \rangle_v$$

$WF_v(A)$ and $SF_v(A)$ are called Fairness together, write as $F_v(A)$.

Defination2.13 TLSA formula:

$$\Phi \triangleq Init_\Phi \wedge \Box [M]_v \wedge F_v(A_1) \wedge \cdots \wedge F_v(A_n)$$

$Init_\Phi$ are initial states of system, M is probability action, A_i（$i \in \mathbb{N}$）is action.

3.2 SEMANTICS

we use symbol "〚 〛" represent semantics, for example, 〚A〛represent the semantics of action A.

According to the definition of actions, it can include the value of variable in nest state, it means that action is a relationship between two state:

$$s[\![A]\!]t \triangleq A(\forall v : s[\![v]\!] / v, t[\![v]\!] / v')$$

State predicate P can regard as action that don't include new value of variable, this means its semantics is irrelevant with next state. Wecan write:

$$s[\![P]\!] \triangleq P(\forall v : s[\![v]\!] / v)$$

Probability action M is disjunction operation with a series of actions under the conditions of probability distribution, its result still is an action:

$$s[\![M]\!]t \triangleq M(\forall v : s[\![v]\!] / v, t[\![v]\!] / v')$$

Defination2.14 *Behaviors* [4]: Behavior is a sequence constituted by unlimited state, marked σ, ith state write as σ_i:

$$\sigma \triangleq \sigma_0 \xrightarrow{M_0} \sigma_1 \xrightarrow{M_1} \sigma_2 \xrightarrow{M_2} \cdots$$

In followed text, we will use $\sigma[..n]$ and σ^{+n} presenting top n state limited sequence or followed unlimited sequence after nth state respectively:

$$\sigma[..n] \triangleq \sigma_0 \xrightarrow{M_0} \sigma_1 \xrightarrow{M_1} \cdots \xrightarrow{M_{n-1}} \sigma_n$$

$$\sigma^{+n} \triangleq \sigma_n \xrightarrow{M_n} \sigma_{n+1} \xrightarrow{M_{n+1}} \sigma_{n+2} \xrightarrow{M_{n+2}} \cdots$$

Of course, σ can be write as $\sigma[..n] \circ \sigma^{+n}$。

To state predicate ***P,*** we have：

$$\sigma[\![P]\!] \triangleq \sigma_0[\![P]\!]$$
$$\sigma^{+n}[\![P]\!] \triangleq \sigma_n[\![P]\!]$$
$$\sigma[\![\Box P]\!] \triangleq \forall n \in N : \sigma^{+n}[\![P]\!]$$

To action ***A***:

$$\sigma[\![A]\!] \triangleq \sigma_0[\![A]\!]\sigma_1$$
$$\sigma^{+n}[\![A]\!] \triangleq \sigma_n[\![A]\!]\sigma_{n+1}$$
$$\sigma[\![\Box A]\!] \triangleq \forall n \in N : \sigma^{+n}[\![A]\!]$$

To probability action ***M***:

$$\sigma[\![M]\!] \triangleq \sigma_0[\![M]\!]\sigma_1$$
$$\sigma^{+n}[\![M]\!] \triangleq \sigma_n[\![M]\!]\sigma_{n+1}$$
$$\sigma[\![\Box M]\!] \triangleq \forall n \in N : \sigma^{+n}[\![M]\!]$$

Similarly, those are the semantics of other concept below:

Eventually: $[\![\Diamond F]\!]$ means formula F can be true eventually. In other words, at present or later, there is a state s: $s \models F$. *In fact ,* $\Diamond F \triangleq \neg\Box\neg F$. Hence it's not hard to get:

$$\sigma[\![\Diamond F]\!] = \exists n \in N : \sigma^{+n}[\![F]\!]$$

***Infinitely often*:** If formula F is true at unlimited states in behavior σ, then we can say $\sigma \models \Box\Diamond F$. In the same, the formal semantics of $\Box\Diamond F$ can be defined below:

$$\sigma[\![\Box\Diamond F]\!] \triangleq \forall n : (\exists m \in N : \sigma^{+n+m}[\![F]\!])$$

***Leads to*:** formula $\Box(F \to \Diamond G)$ is true, if and only if that F is true must lead to G is true at later state. Its formal semantics is:

$$\sigma[\![\Box(F \to \Diamond G)]\!] \triangleq \forall n : (\sigma^{+n}[\![F]\!] \Rightarrow (\exists m \in N : \sigma^{+(n+m)}[\![G]\!]))$$

3.3 Probabilistic property

Besides common qualitative properties of concurrent system, such as *invariance*, *eventuality* [4], we also analysis the quantitative properties of probabilistic system.

The probabilistic system defined in defination1.1 is a label transition system with probability distribution of action, so we can specify and verify some probabilistic properties of system actions. For example, after action A occurring, the probability of action B occur eventually is not less than 0.4, that is $\mathbb{P}_{\geq 0.4}(A \to \Diamond B)$. Those properties are mainly used to describe the reliability and performance of systems.

IV. SPESIFYING PROBABILISTIC SYSTEM BY TLSA

We observe a biological group, use x representing the quantity of that group, supposing the biggest quantity is 5, at next moment the quantity may be three kind variation below: increasing 1, reducing 1, not change; and besides extreme states of x=0 and x=5, the probability of three kind variation is 1/3; at state of x=0, the probability of no change is 1; at the state of x=5, the probability of reducing 1 is 1. Its probabilistic transition graph is as figure 1.

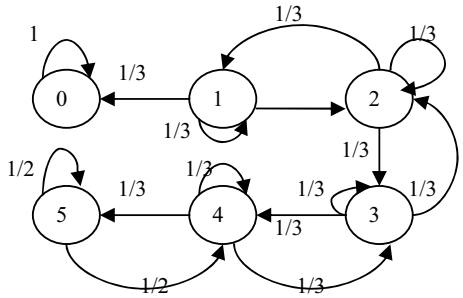


Figure 1. Biological group quantity transition graph

Figure 1 indicate a classical probabilistic system, its state space $\mathcal{S} \triangleq \{x \mid x = 0,1,2,3,4,5\}$, $Init_\Phi \triangleq x = 3$, three actions are: $A \triangleq x' = x+1$, $B \triangleq x' = x-1$ and $C \triangleq x' = x$; probability distribution of actions is: at state 1, 2, 3, 4, the probability of three actions occurring is equal to 1/3; at state 0, the probability of action C is 1; at state 5, the probability of action B and C are all 1/2.

We specify the probabilistic system as figure 1 using TLSA below:

----------MODULE GenExt----------
EXTENDS Naturals, Reals
VARIABLE x
$GEini \triangleq x = 3$
$GEnxtA \triangleq x' = x+1$
$GEnxtB \triangleq x' = x-1$
$GEnxtC \triangleq x' = x$
IF $s \neq 0, s \neq 5, a = A,B,C$ THEN $p(s,a) \triangleq 1/3$
IF $s = 5, a = B,C$ THEN $p(s,a) \triangleq 1/2$
IF $s = 0, a = C$ THEN $p(s,a) \triangleq 1$
$GEnxt \overset{p(\cdot)}{=} \vee GEnxtA$
$\vee GEnxtB$
$\vee GEnxtC$
$GE \triangleq GEini \wedge \Box[GEnxt]_x \wedge WF_x(GEnxtA)$

V. MODEL CHECKING OF PROBABILISTIC PROPERTIES

If the system as figure 1 satisfy that the probability of action A occurring lead to action B occurring is 0.5, that is $\mathbb{P}_{\leq 0.5}(A \rightarrow\!\!\!\!\!\!\diamond\; B)$.

We can embed probability module in system specification by TLSA.

----------MODULE GenExt----------
EXTENDS Naturals, Reals
VARIABLE x
CONSTANT Rpr
ASUME $Rpr \in [0,1]$

----------MODULE Prob----------
VARABLE pr

$Pini \triangleq pr = 0$

$Pnxt \triangleq pr' = IF\ GEnxtB\ THEN\ p(\cdot, GEnxtB)$
$ELSE\ IF\ pr = 0\ THEN\ 1 - p(\cdot, GEnxtB)$
$ELSE\ pr * (1 - p(\cdot, GEnxtB))$

$GEProb \triangleq Pini \wedge \Box[Pnxt]_{pr}$

$P(pr) \triangleq INSTANCE\ Prob$

$GEini \triangleq x = 3$

$GEnxtA \triangleq x' = x + 1$

$GEnxtB \triangleq x' = x - 1$

$GEnxtC \triangleq x' = x$

IF $s \neq 0, s \neq 5, a = A, B, C$ THEN $p(s,a) \triangleq 1/3$

IF $s = 5, a = B, C$ THEN $p(s,a) \triangleq 1/2$

IF $s = 0, a = C$ THEN $p(s,a) \triangleq 1$

$GEnxt \overset{p(\cdot)}{=} \vee GEnxtA$
$\vee GEnxtB$
$\vee GEnxtC$

$GE \triangleq GEini \wedge \Box[GEnxt \wedge P(pr)!Pnxt]_x \wedge WF_x(GEnxtA)$

THEOREM $GE \Rightarrow \mathbb{P}_{\leq 0.5}(GEnxtA \rightarrow \Diamond GEnxtB)$

The model checking tools of TLSA can be obtained by adding probabilistic logic operator and definition of probability calculation module, we don't discuss it here.

VI. CONCLUSION

Based on Temporal Logic of Actions, we put forward a Temporal Logic of Stochastic Actions (TLSA), by adding state action probability attribution and probabilistic operator; it is fit for specifying and model checking probabilistic systems. Compared with Probabilistic Computation Tree Logic (PTCL)[2], Continuous Stochastic Logic (CSL)[3] and Real Timed Probabilistic Computation Tree Logic(RPTCL)[7], the most important feature of TLSA is that it can describe system , and it can describe the properties of system in same time. This is useful for verification of system properties and refinement of system design.

In addition, the probability distribution is based on system actions in TLSA, and the quantity of actions are fewer than system state, this is beneficial to specifying and calculating of probabilistic properties.

The model checking tools of TLSA can be obtained by extending TLC.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (NO. 61163001/F020106), Natural Science Foundation of Guizhou Province (NO. J [2012]2096), and High-level Talent Recruitment Foundation of Guizhou University of finance and Economics.

REFERENCES

[1] Katoen, J.P. Perspectives in probabilistic verification. Proceedings 2nd IEEE/IFIP International Symposium on Theoretical Aspects of Software Engineering, Nanjing, China, 2008.

[2] C. Baier and J.-P. Katoen. Principles of Model Checking. MIT Press, 1st edition, 2008.

[3] Baier, Christel and Cloth, Lucia and Haverkort, Boudewijn R. and Kuntz, Matthias and Siegle, Markus. Model Checking Markov Chains with Actions and State Labels. IEEE Transactions on Software Engineering, 33 (4), 2007. pp. 209-224.

[4] Leslie Lamport. The temporal logic of actions. ACM Transactions on Programming Languages and Systems (TOPLAS), v.16 n.3, p.872-923, 1994.

[5] Luca De Alfaro , Zohar Manna, Formal verification of probabilistic systems, Stanford University, Stanford, CA, 1998.

[6] A. Kucera, J. Esparza, R. Mayr. Model checking probabilistic pushdown automata. LMCS 2(1): (2006).

[7] M.Z. Kwiatkowska, G. Norman, R. Segala, J. Sproston: Automatic verification of real-time systems with discrete probability distributions. Theor. Comput. Sci. 282(1): 101-150, 2002.

Quantum-behaved Flower Pollination Algorithm

Kezhong Lu
Department of Computer Science
Chizhou University
Chizhou, China
e-mail: luke76@163.com

Haibo Li
Institute of things engineering
Wuxi Institute of Commerce
Wuxi, China
e-mail: li780717@163.com

***Abstract*—Flower Pollination algorithm (FPA) is a new nature-inspired algorithm, based on the characteristics of flowering plants. In this paper, a new hybrid optimization method called quantum-behaved flower pollination algorithm (QFPA) is proposed. The method combines the standard Flower Pollination algorithm (FPA) with the quantum- behaved search mechanism to improve the global searching ability and accuracy. The simulation experiments show that the proposed QFPA algorithm can improve the convergence speed and the quality of solutions effectively.**

Keywords-flower pollination algorithm; quantum-behaved search mechanism; optimization

I. I INTRODUCTION

In recent years, more and more bioinspired algorithms are proposed, such as particle swarm optimization (PSO), bacterial foraging optimization algorithm (BFOA), invasive weed optimization (IWO), harmony search (HS), firefly algorithm(FA) and bat algorithm(BA) et al. Because of their advantages of global, parallel efficiency and universality, swarm intelligence algorithms have been widely used in engineering optimization, scientific computing, automatic control, and other fields[1].

Flower pollination algorithm (FPA) developed in 2012 by Xin-She Yang is inspired by flower pollination process of flowering plants in nature[2]. Preliminary studies suggest that the FPA can perform superiorly, compared with genetic algorithm and particle swarm optimization, and it is applicable for mixed variable and engineering optimization, such as text clustering[3], wireless sensor network[4], integer programming problems[5], sudoku puzzles [6], et al.

Some improved FPA algorithms are proposed in the last three years by scholars. For solving multidimension function optimization problems, the whole update and evaluation strategy on solutions may deteriorate the convergence speed and the quality of solution of algorithm due to interference phenomena among dimensions. To overcome this shortage, R. Wang proposed a dimension by dimension improvement [1]. O. Abdel-Raouf proposed a new hybrid optimization method called hybrid flower pollination algorithm. The method combines the standard flower pollination algorithm with the particle swarm optimization algorithm to improve the searching accuracy[7]. H. Xiao proposed a hybrid algorithm of simulated annealing and flower pollination algorithm to overcome the problems of low accuracy computation, slow speed convergence and being easily relapsed into local extremum[8]. To improve the FPA algorithm searching accuracy, O. Abdel-Raouf combines the standard Flower Pollination algorithm (FPA) with the chaotic Harmony Search (HS) algorithm[6].

To overcome local convergence, we introduce quantum mechanics into FPA algorithm in this pape. The new-FPA is named as quantum-behaved flower pollination algorithm (QFPA). The rest of this paper is organized as follows: it outlines the flower pollination algorithm in section II, and then new flower pollination algorithm is described in section III. Experimental simulation and analysis are presented in section IV. Section V concludes the paper.

II. FLOWER POLLINATION ALGORITHM

Flowering plants flow pollination process inspired Xin-She Yang to develop Flower Pollination Algorithm (FPA) in 2012. For ease, the four rules given below are used[2].

Rule 1: Biotic and cross-pollination is considered as global pollination process with pollen-carrying pollinators performing Lévy flights.

Rule 2: Abiotic and self-pollination are considered as local pollination.

Rule 3: Flower constancy can be considered as the reproduction probability is proportional to the similarity of two flowers involved.

Rule 4: Local pollination and global pollination is controlled by a switch probability $p \in [0, 1]$, with a slight bias toward local pollination.

Due to the physical proximity and other factors such as wind, local pollination can have a significant fraction p in the overall pollination activities. In the global pollination step, flower pollens are carried by pollinators such as insects, and pollens can travel over a long distance. This ensures the pollination and reproduction of the most fittest, and thus we represent the most fittest as g_*. The first rule plus flower constancy can be represented mathematically as (1).

$$x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t). \quad (1)$$

Where x_i^t is the solution vector i at iteration t, g_* is the current best solution, γ is a scaling factor to control the step size, and L is the strength of the pollination which is a step

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.24

size randomly drawn from Lévy distribution. We draw $L > 0$ from a Lévy distribution:

$$L \sim \frac{\lambda\Gamma(\lambda)\sin(\pi\lambda/2)}{\pi}\frac{1}{s^{1+\lambda}}, (s>>\sigma_0>0). \quad (2)$$

Here $\Gamma(\lambda)$ is the standard gamma function, and this distribution is valid for large steps $s > 0$. In theory, it is required that $|s_0|>>0$, but in practice s_0 can be as small as 0.1. In all our simulations below, we have used $\lambda = 1.5$.

The local pollination (Rule 2) and flower constancy can be represented as

$$x_i^{t+1} = x_i^t + \varepsilon(x_j^t - x_k^t). \quad (3)$$

Where x_j^t and x_k^t are solution vectors drawn randomly from the solution set. The parameter ε is drawn from uniform distribution in the range from 0 to 1. Though Flower pollination activities can occur at all scales, both local and global, adjacent flower patches or flowers in the not-so-far-away neighborhood are more likely to be pollinated by local flower pollen than those faraway. In order to imitate this, we can effectively use the switch probability like in Rule 4 or the proximity probability p to switch between common global pollination and intensive local pollination. A preliminary parametric showed that p=0.8 might work better for most applications[2].

III. QUANTUM-BEHAVED FLOWER POLLINATION ALGORITHM

A. Quantum-behaved Search Mechanism

Sun et al. [9]assumed that a PSO system is a quantum system, each particle has a quantum behavior with its quantum state formulated by a wave function ψ, and proposed a quantum-behaved search mechanism. Assumed that, at iteration t, particle i moves in d-dimensional space with a δ potential well centered at $p_{i,j}^t$ on the jth dimension. Correspondingly, the wave function at iteration t+1 is

$$\psi(x_{i,j}^{t+1}) = \frac{1}{\sqrt{\mathrm{H}_{i,j}^t}}\exp(-|x_{i,j}^t - p_{i,j}^t| / \mathrm{H}_{i,j}^t). \quad (4)$$

Where $\mathrm{H}_{i,j}^t$ is the standard deviation of the double exponential distribution, varying with iteration number t. Hence, the probability density function Q is a double exponential distribution as follows

$$Q(x_{i,j}^{t+1}) = |\psi(x_{i,j}^t)|^2 = \frac{1}{\mathrm{H}_{i,j}^t}\exp(-2|x_{i,j}^t - p_{i,j}^t| / \mathrm{H}_{i,j}^t). \quad (5)$$

and thus the probability distribution function F is

$$F(x_{i,j}^{t+1}) = 1 - \exp(-2|x_{i,j}^t - p_{i,j}^t| / \mathrm{H}_{i,j}^t). \quad (6)$$

Using Monte Carlo method, we can obtain the jth component of position x_i at iteration t+1 as

$$x_{i,j}^{t+1} = p_{i,j}^t \pm \frac{1}{2}\mathrm{H}_{i,j}^t \ln(1/u_{i,j}^t). \quad (7)$$

where $u_{i,j}^t$ is a random number uniformly distributed over (0, 1). The value of $\mathrm{H}_{i,j}^t$ is calculated as

$$\mathrm{H}_{i,j}^t = 2\alpha|c_j^t - x_{i,j}^t|. \quad (8)$$

where c^t is known as the mean best (mbest) position defined as the mean of the positions of all particles. That is

$$c^t = (c_1^t, c_1^t, ..., c_d^t) = (\frac{1}{M}\sum_{i=1}^{M} x_{i,1}^t, \frac{1}{M}\sum_{i=1}^{M} x_{i,2}^t, ..., \frac{1}{M}\sum_{i=1}^{M} x_{i,d}^t). \quad (9)$$

where M is the population size and x_i^t is the personal position of particle i. Hence, the position of the particle updates according to the following equation

$$x_{i,j}^{t+1} = p_{i,j}^t \pm \alpha|c_j^t - x_{i,j}^t| \ln(1/u_{i,j}^t). \quad (10)$$

where parameter α is known as the contraction–expansion coefficient, which can be tuned to control the convergence speed of the algorithms. it is suggested that decreasing the value of a linearly from 1.0 to 0.5[10]. The value of α is computed by

$$\alpha = 1 - t/T * 0.5. \quad (11)$$

where T is the maximum number of iterations.

B. Flower Pollination Algorithm with Quantum-behaved Search Mechanism

In quantum-behaved search mechanism, the particle is assumed to have quantum behavior and its position and velocity of a particle cannot be determined simultaneously according to uncertainty principle. Therefore, the quantum-behaved particle can fly more randomly in searching space. The quantum-behaved search mechanism enhances the global search ability of the PSO algorithm[9]. Here, we introduce quantum-behaved search mechanism into flower pollination algorithm, to enhances the global search ability of the FPA algorithm.

The main search process in FPA algorithm is the global pollination process by (1). Supposed that the p_i^t in (10) is generated by (1), we substitute (1) into (9) and get (12).

$$x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t) \pm \alpha \,|\, c^t - x_i^t \,|\, \ln(1/u_i^t) \,. \quad (12)$$

So the solution vector *i* updated by (12) has the quantum properties, it is favorable to enhance the global convergence ability. The flower pollination algorithm updated solution vector with (12) instead of (1) is called quantum-behaved flower pollination algorithm (QFPA).

C. Steps of QFPA Algorithm

The QFPA algorithm is described in fig.1.

Define objective function $f(x)$, $x = (x_1, x_2, ..., x_d)$
Initialize a population of M flowers/pollen gametes
Find the best solution g_* in the initial population
Define a switch probability p [0, 1] and scaling factor γ
while (t <T)
  $c^t = (\frac{1}{M}\sum_{i=1}^{M} p_{i,1}^t, \frac{1}{M}\sum_{i=1}^{M} p_{i,2}^t, ..., \frac{1}{M}\sum_{i=1}^{M} p_{i,d}^t)$
  for i=1:M
    if (rand(0,1)<p)
      u=rand(0,1)
      α=1-t/T*0.5
      Draw a (d-dimensional) step vector L which obeys a Lévy distribution
      if rand(0,1)>0.5
        $x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t) + \alpha \,|\, c^t - x_i^t \,|\, \ln(1/u_i^t)$
      else
        $x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t) - \alpha \,|\, c^t - x_i^t \,|\, \ln(1/u_i^t)$
      end if
    else
      ε=rand(0,1)
      $x_i^{t+1} = x_i^t + \varepsilon(x_j^t - x_k^t)$
    end if
    Evaluate new solutions
    If new solutions are better, update them
  end for
  Find the current best solution g_*
end while

Figure 1. Pseudo code of the QFPA algorithm.

IV. SIMULATION AND ANALYSIS

A. Benchmark Functions

To validate our proposed quantum-behaved flower pollination algorithm(QFPA). We have used 8 benchmark functions, including four multimodal functions f_1-f_4 and four unimodal functions f_5-f_8. The analytical form each function, along with their names, and bounds of search space are shown in Tab.1. The global minimum values of the eight benchmark functions are all 0.

B. Parameter Settings

The FPA and QFPA algorithms are tested with 30 independent runs on each of the test functions listed in Table 1. The swarm size is set to 50 for both algorithms. The number of generations T=1500, the dimention d= 30, and the switching probability p=0.8. The scaling factor (γ) changes with the problems, and it is set to 0.1, 1, 0.5, 0.1, 15, 5, 0.1 and 25 for functions f_1 to f_8 respectively. The α in QFPA algorithm is caculated as (11). We have used Matlab version R14 for the simulation.

TABLE I. BENCHMARK FUNCTIONS USED IN THE EXPERIMENTAL STUDIES. HERE, S: SEARCH SPACE.

No.	Name	S	Function Definition
f_1	Griewank	[-600, 600]	$f_1 = \frac{1}{4000}\sum_{i=1}^{d} x_i^2 - \prod_{i=1}^{d} \cos(x_i/\sqrt{i}) + 1$
f_2	Schwefel	[-500, 500]	$f_2 = 418.9829d - \sum_{i=1}^{d} x_i \sin(\|x_i\|^{1/2})$
f_3	Ackley	[-32, 32]	$f_3 = -20e^{-0.2\sqrt{\sum_{i=1}^{d} x_i^2/d}} - e^{\sum_{i=1}^{d}\cos(2\pi x_i)/d} + 20 + e$
f_4	Rastrigin	[-15, 15]	$f_4 = \sum_{i=1}^{d} (x_i^2 - 10\cos 2\pi x_i + 10)$
f_5	Sphere	[-5.12,5.12]	$f_5 = \sum_{i=1}^{d} x_i^2$
f_6	Zakharov	[-5, 10]	$f_6 = \sum_{i=1}^{d} x_i^2 + (\sum_{i=1}^{d} 0.5ix_i)^2 + (\sum_{i=1}^{d} 0.5ix_i)^4$
f_7	Rosenbrock	[15,15]	$f_7 = \sum_{i=1}^{d} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$
f_8	Quartic	[-1.28,1.28]	$f_8 = \sum_{i=1}^{d} ix_i^4$

C. Experimental Results

In FPA, the global and local pollination technique is used to balance between explorations and exploitations. In global pollination, the Lévy distribution is applied to generate new solutions. Because the Lévy distribution generates new solutions with small mutation step size, the algorithm is unable to explore the whole search space and is likely to fail to escape from the intermediate locally optimal. In QFPA a quantum-behaved search mechanism is introduced. In quantum space, QFPA has the capability to generate new solutions with bigger mutation step size. Thus the QFPA algorithm is more likely to escape from the locally optimal points. From the experimental results, we can see that the QFPA algorithm show much improved results than the FPA algorithm on the continuous optimization problems.

For simulation, eight benchmark functions have been used in this experiment. Among them four functions are unimodal and four are multimodal. A unimodal function is one that has a single local minimum while a multimodal function is a function with many local minima. In Table 2, we can see that the solution quality (mean value) is much better for QFPA than FPA on both multimodal and unimodal functions. Also the value of the standard deviation indicates that the result of FPA is more consistent. The convergence characteristics of FPA and QFPA on the

multimodal Schwefel function and the unimodal Sphere function are plotted in Fig.2 and Fig.3 respectively, which show the superior convergence behavior of QFPA.

TABLE II. COMPARISON BETWEEN FPA AND QFPA ON 8 STANDARD BENCHMARK FUNCTIONS.

Fun	Algorithm	Best	Worst	Mean	SD
f_1	FPA	1.22e-3	2.53e-2	6.91e-3	6.98e-3
	QFPA	3.87e-8	4.73e-5	1.27e-5	1.28e-5
f_2	FPA	3.48e2	4.44e3	4.11e3	2.51e2
	QFPA	3.01	1.88e2	3.02e1	3.67e1
f_3	FPA	3.37e-4	1.71e-3	7.06e-4	2.63e-4
	QFPA	5.78e-9	1.12e-4	8.22e-6	2.42e-5
f_4	FPA	7.78e1	1.36e2	1.08e2	1.12e1
	QFPA	4.34e-12	3.01e-3	1.72e-4	5.68e-4
f_5	FPA	6.71e-8	3.04e-7	1.83e-7	7.09e-8
	QFPA	2.38e-19	8.83e-15	4.45e-16	1.61e-15
f_6	FPA	1.34e1	3.65e1	2.41e1	6.51
	QFPA	4.00e-2	1.36e1	2.28	3.56
f_7	FPA	3.42e1	1.09e2	5.28e1	2.65e1
	QFPA	2.84e1	2.93e1	2.89e1	0.22
f_8	FPA	1.19e-13	2.69e-12	8.70e-13	7.30e-13
	QFPA	2.72e-17	1.00e-13	1.50e-14	2.69e-14

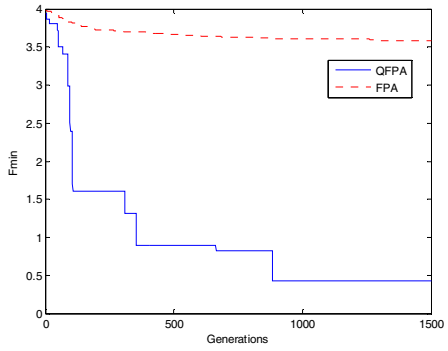


Figure 2. Convergence characteristics of FPA and QFPA on f_2.

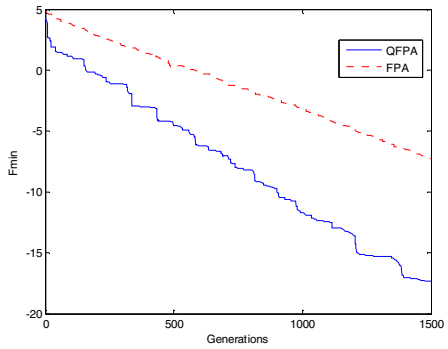


Figure 3. Convergence characteristics of FPA and QFPA on f_5.

V. CONCLUSION

In this paper, a variant of FPA, namely QFPA, is proposed by using a quantum mechanism in the global pollination process. By eight typical standard benchmark functions simulation the results show that QFPA algorithm generally has strong global searching ability, and effectively avoid the defects of FPA algorithms fall into local optimization. QFPA has improved the convergence speed and convergence precision of FPA. The experiment results show that it is an effective algorithmto solve complex functions optimization problems. In this paper, we only consider the global optimization. The algorithm can be extended to solve other problems such as constrained optimization problems and multiobjective optimization problem. In addition, many engineering design problems are typically difficult to solve. The application of the proposed QFPA algorithm in engineering design optimization may prove fruitful.

REFERENCES

[1] R. Wang and Y.Q. Zhou, "Flower Pollination Algorithm with Dimension by Dimension Improvement," Mathematical Problems in Engineering, vol. 2014, pp. 1–9,2014.

[2] Y. Xin-She, "Flower pollination algorithm for global optimization," Unconventional Computation and Natural Computation, Lecture Notes in Computer Science, vol. 7445, pp. 240–249, 2012.

[3] M. Kaur and N. Kaur, "Text Clustering using PBO algorithm for Analysis and Optimization", International Journal of Current Engineering and Technology, vol.4, no.6, pp.3876-3878,2014.

[4] M. Sharawi, E. Emary, I. A. Saroit, et al., "Flower pollination optimization algorithm for wireless sensor network lifetime global optimization,"International Journal of Soft Computing and Engineering, vol. 4, no. 3, pp. 54–59, 2014.

[5] I. El-Henawy and M. Ismail, "An improved chaotic flower pollination algorithm for solving large integer programming problems," International Journal of Digital Content Technology and Its Applications, vol. 8, no. 3, pp. 72–81, 2014.

[6] O. Abdel Raouf, I. El-henawy, and M. Abdel-Baset, "A novel hybrid flower pollination algorithm with chaotic harmony search for solving sudoku puzzles," International Journal of Modern Education and Computer Science, vol. 3, pp. 38–44, 2014.

[7] O. Abdel-Raouf, M. Abdel-Baset, and I. El-henawy, "A New Hybrid Flower Pollination Algorithm for Solving Constrained Global Optimization Problems", International Journal of Applied Operational Research, vol. 4, no. 2, pp. 1-13, 2014.

[8] H. Xiao, C. Wan, Y. Duan, et al., "Flower pollination algorithm based on simulated annealing", Journal of Computer Applications, vol. 35, no. 4, pp.1062-1066,1070,2015.

[9] J. Sun, B. Feng,and W. Xu,"A global search strategy of quantum-behaved particle swarm optimization," IEEE Conference on Cybernetics and Intelligent Systems. Piscataway, NJ: IEEE Press, pp. 111-116,2004.

[10] H. Long, W. Xu, X. Wang, et al., "Using Selection to Improve Quantum-behaved Particle Swarm Optimization," Control and Decision, vol.25,no.10,pp.1499-1506, 2010.

Multi-objective flexible job shop schedule based on ant colony algorithm

Jiang Xuesong
Qilu University Of Technology
Ji Nan, China
jxs@qlu.edu.cn

Tao Qiaoyun
Qilu University Of Technology
Ji Nan, China
785484360@qq.com

***Abstract*—In this paper, an improved ant colony algorithm is proposed to solve solving multi-objective flexible shop scheduling problem. Limitations of the traditional ant colony algorithm weighting coefficient method will result in a greater impact on the results because the determination of the weighting factor has greater subjective factors. Proposed algorithm adds a set of BPs to save all the Pareto set ant appear after iteration, the algorithm improves the search capabilities of the ant colony. The convergence speed is improved on ameliorating the pheromone update rule based on the global optimal experience to guide the optimization way .Thus, multi-objective Flexible Job Shop Scheduling Problems Pareto optimal solution was conducted. Finally, the proposed theory in this paper is proved to solve the multi-objective flexible job shop scheduling optimization problems by examples.**

Keywords-Multi-objective optimization; Flexible job shop scheduling problems; Ant colony algorithm; Pareto optimal solution;

I. INTRODUCTION

Flexible job shop scheduling problem is an extension of the classical job-shop scheduling problem [1]. The data indicate that 95 percent of the time in the manufacturing process was consumed in the non-cutting process [2]. Therefore, the study of multi-objective flexible job shop scheduling problem has important theoretical and engineering significance [3].

The flexible job shop scheduling problem recently captured the interests in many researchers. Bruker and Schlie [4] were the first to consider this problem by developing a polynomial algorithm to solve the FJSP with two jobs. Motaghedi et al [5] presented an effective hybrid genetic algorithm to solve the multi-objective FJSP. Zhang et al [6] proposed a GA with tabu search procedure for FJSP with transportation constraints and bounded processing times to minimize the makespan and the storage of solutions. General Pan et al [7] designed a two-way production workshop scheduling method using cycle and key artifacts delivery time for production as optimization goal. Zhang Wei et al [8]such as the use of the main store, handle the relationship between this problem in each target from hierarchical structures colonies and particle swarm optimization.

In this paper, an improved ant colony algorithm is proposed to solve multi-objective flexible shop scheduling problem. A multi-objective scheduling model is given. To improve the search capabilities, we add an external set of BP(t) to save all the Pareto disaggregation after iteratio in the ant colony algorithm. By the way of the global optimum experience-based guidance, all the non-dominated solutions that we currently found are preserved, and then use these solutions to guide the optimization area ant to accelerate the convergence rate. Finally, simulation experiments proved that the algorithm can make the production system meet the target of reasonable configuration and effective utilization of resources with the highest efficiency and lowest cost under certain constraints.

II. MULTI-OBJECTIVE FLEXIBLE JOB SHOP SCHEDULING

A. Problem Description

The flexible job-shop scheduling considers n jobs to be processed on m machines, where each job i consists of a sequence of n_i operations V_{ij} ,j=1, 2, ⋯, n_i . For the flexible job-shop scheduling, it needs to determine both the assignment of machines and the sequence of operations on all the machines to optimize multiple scheduling objectives under the conditions meet the constraints. In this paper, the following three objectives which include the minimum processing cost, the shortest processing time and the highest rate of qualified products are to be optimized. Mathematically, the corresponding optimization model is described as follows:

$$f_1(\mathrm{x}) = \min \sum_{i=1}^{N} \sum_{j=1}^{n_i} x_{ijk} t_{ijk} \tag{1}$$

$$\begin{aligned} f_2(\mathrm{x}) &= \min \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{n_i} c_{ijk} x_{ijk} \\ &= \min \sum_{m=1}^{M} \sum_{i=1}^{N} (\sum_{j=1}^{n_i} E_{ijk} x_{ijk} + \sum_{j=1}^{n_i} V_{ijk} x_{ijk}) \end{aligned} \tag{2}$$

$$f_3(\mathrm{x}) = \max \sum_{m=1}^{M} \sum_{i=1}^{N} h_{ij} \tag{3}$$

$$\min/\max f(x) = F(x) = (f_1(x), f_2(x), f_3(x)) \tag{4}$$

$$\sum_{k=1}^{M} S_{ijk} x_{ijk} \geq \sum_{k=1}^{M} \left[(\mathrm{S}_{i(\mathrm{j}-1)\mathrm{k}} \mathrm{t}_{i(\mathrm{j}-1)\mathrm{k}}) \right] x_{i(\mathrm{j}-1)\mathrm{k}} \tag{5}$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.25

$$X_{ijk} = \begin{cases} 1, & if\ V_{ij}\ is\ done\ on\ machine\ k \\ 0, & elsewhere \end{cases} \quad (6)$$

$$R_{ijmnq} \begin{cases} 1, & if\ V_{ij}\ is\ done\ on\ machine\ i\ in\ priority\ m \\ 0, & elsewhere \end{cases} \quad (7)$$

We use the following notations:

n :Number of jobs ; m :Number of machines; V_{ij} :The $_j$ operation of job i ; X_{ijk} :The operation $_j$ of job i is assigned to machines k ; t_{ijk} :Completion time of the operation $_j$ in job i by machine k ; R_{ijmnq} :The sequence between operation $_j$ on the machine i and the processes n of machining m operation; C_{ijk} :the i step of the operation on the $_j$ path of the machine k processing costs; S_{ijk} :Start time of operation V_{ij} on machine k ; E_{ijk} :The machine cost of operation V_{ij} on machine k; h_{ij} :Product performance indicators ;

In the Flexible Job Shop Scheduling Problem, some assumptions are made as follows:

- All jobs can be started at time 0.
- All machines are available at time 0.
- Each machine can process only one job at a time.
- Each job can be processed by only one machine at a time.
- Each machine cannot be interrupted before it finishes the job's work.
- The processing of an operation cannot be interrupted once started.
- Machine break down does not occur, which means all the machines are continuously available throughout the production stage.
- Job transportation time among machines is not considered.

B. Optimization Algorithm

For a weighting factor based on ant colony algorithm for solving multi-objective flexible method shop scheduling method limitations, this paper presents an improved ant colony algorithm for solving this problem. In order to guide the ants search for viable space efficient, adding an external set of $BP(t)$ which is used to save all Pareto sets the entire ant colony after the t iterations found. We are looking for walking the most sparse solutions in the set of $BP(t)$. To find the most sparsely walking non-dominated solutions in the collection $BP(t)$ .Assuming the current collection $BP(t)$ in a non-dominated solutions q $x=(x_1,x_2,x_3,\ldots,x_q)$, if ant i enters the collection $BP(t)$, indicating that the location is not dominated by ants, ant i as optimization direction conducive algorithm direction toward the Pareto frontier evolution. Therefore, it should increase the ant i pheromone to guide other ants in the area of the location where the ant i search; on the contrary, it should be appropriate to reduce the pheromone. Then, by optimizing the way the global optimum experience based guidance, the current discovery of all non-dominated solutions saved, and then use these solutions to guide the other ant optimization area, improving the convergence rate.

C. Global optimum experience guided optimization approach

Due to constraints on multiple goals are mutually exclusive, and some of the goals of improvement will always cause deterioration of some goals, therefore, flexible shop scheduling multi-objective optimization problem, each target can't reach their optimum value. For multi-objective optimization problem, the merit of its solution is relative, there is no absolute optimum. We can only get a set of optimal solutions set, the set of solutions cannot further comparison between the advantages and disadvantages of each other, so that the solution set of generally called Pareto optimal solution set [9]. Due to the Pareto optimal solution set of points can be used as a potential solution, so we requirement non inferiors set to approximate real Pareto solution set of asking questions, and get the Pareto frontier.

In this algorithm, we added a collection of external $BP(t)$ .To distinguish each ant pheromone increment, when multiple ants enter into the collection BP at the same time, we put the new ant i enters into the collection $BP(t)$ and the original collection $BP(t)$ of the solution to the objective function value of minimum distance $\theta(t)$ as the location of the ant i released pheromone.

$$\theta(t) = \min \sqrt{\sum_{i=1}^{n} (f_i(x) - f_i(x_v))^2} \quad x_v \in BP(t) \quad (8)$$

Therefore, the ant colony algorithm pheromone update is defined as follows:

$$\tau_i(t+1) = \begin{cases} \rho\tau_i(t) + \theta(t), x \in BP(t+1) \\ \rho\tau_i(t), Otherwise \end{cases} \quad (9)$$

D. State transition rule

Simply put, the state transition rules are the rules to select the next node. In multi-objective optimization algorithm, the ant i movement is associated with the element information and the distance of the ant j . High and closer ants pheromone concentration should be at a

higher probability is selected as the next step in the direction of the movement direction. See below transition probability formula:

$$P_{ij}(\mathrm{t})=\begin{cases}\dfrac{\left[\tau_{ij}(\mathrm{t})\right]^{\alpha}(\eta_{ij})^{-\beta d_{ij}}}{\sum_{s\in J_k}\left[\tau_{is}(\mathrm{t})\right]^{\alpha}(\eta_{is})^{-\beta d_{ij}}}, (d_{ij}<0\cap i\neq j)\cup(d_{ij}=0\cap i=1) \\ 0, Otherwise\end{cases} \quad (10)$$

Coefficient α indicates that the amount of information on the extent and Coefficient β indicates that degree of attention node valued by the heuristic information. η_{ij} was on behalf of the visibility of information, take the reciprocal of the objective function increment in the solution process remains unchanged. Probability formula make ensure that the probability of the ant i by moving to other areas to get a better solution ant j, or stay in place.

E. Algorithm implementation steps

Step 1: Initialization algorithm parameters. Include: the number of ants k, the number of iterations t, initial pheromone; ants will want to visit the set of all processes GK, the next step to allow access to the process set SK, each ant has gone through a process set JK. Individual ants are randomly assigned to the GK. Initialize the collection $BP(\mathrm{t})$ t=0;

Step 2: Choose search path. Ant k (k = 1，2，3 ..., k) in accordance with the transition probability equation (10) to select the next step in the collection to arrive in SK.

Step 3: Update the collection process. Update Collection $BP(\mathrm{t})$. After ants k chose a process based on step two, put it to add to the collection JK taboo table, deleting the procedure from the set SK and collections GK, update set SK, if this processes is not carried out the final steps, then the subsequent process is added to the SK, and repeat the process until the set GK is empty.

Step 4: Update pheromone. Using Equation (9) updates pheromone.

Step 5: Repeat step two to step five, next-generation search of ants to find the global optimum and iterative optimal ants, until the termination condition is met.

III. EXAMPLES OF APPLICATION

A. Example Overview

Taking a Heavy Machinery Group of lifting equipment paint shop, we are multi-objective optimization workshop scheduling. The main equipment for the coating processing workshop, including pretreatment device, cathode electrophoresis equipment, drying machine, polishing machine, glue manipulator and online testing equipment, etc. In the workshop of a state of the operation process of the production line and record data and order list, we obtain the parameters such as T_{ijk}, C_{ijk} and h_i etc. After computing, we get six process has several process 10 sets of equipment, each artifact of flexible shop scheduling problem. Specific piece of information as shown in Table 1 machine [10], in which that the number 1 was representative of occupancy and the number 0 means no. Index values for different devices and processing time and processing costs of the different processes in Table 2[10].

TABLE I. RELATION BETWEEN MACHINE AND OPERATION

Processes / Equipment	Stepping electrophoresis		Spraying equipment		Machining Center		Drying Equipment			Grinding machine
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Paint pretreatment	1	1	0	0	0	0	1	1	1	0
Primer treatment	0	0	1	1	0	0	1	1	1	0
Putty	0	0	0	0	1	1	0	0	0	0
Polished	0	0	0	0	0	0	0	0	0	1
Intermediate coat paint	0	0	1	0	0	0	0	1	1	0
Topcoats	0	0	0	1	0	0	1	0	1	0
Check up	0	0	0	0	1	0	0	0	0	0

TABLE II. MACHINING PROCESSES INDEX VALUES

Step	numbe r	Piece 1		Piece2		Piece3		Piece 4		Piece5		Piece6	
		Time	Costs	Time	Costs	Time	Costs	Time	Costs	Time	Costs	Time	Costs
Paint pretreatmen t	M1	45	7	14	25	40	21	40	34	4	10	48	39
	M2	74	71	28	3	48	70	47	99	88	46	14	35
	M7	3	12	93	6	19	16	45	77	7	97	18	94
	M8	79	76	60	48	10	66	25	80	48	45	68	22
	M9	64	70	10	10	74	99	87	15	58	30	12	18
Primer treatment	M3	23	73	27	32	90	96	42	16	22	40	14	6
	M 4	31	36	56	29	88	35	24	7	9	65	56	64
	M7	53	71	76	34	29	33	37	48	26	83	74	22
	M8	43	58	95	7	42	3	10	51	68	90	94	78
	M9	79	62	29	59	4	76	46	83	8	22	5	90
Putty	M5	23	35	18	43	67	33	60	49	12	42	33	40
	M6	78	36	19	23	48	13	25	19	78	24	13	12
Polished	M10	49	3	23	12	80	27	60	11	36	20	13	21
Intermediate coat paint	M3	54	28	99	54	58	21	36	21	55	75	19	56
	M8	28	73	43	28	79	90	46	59	41	97	94	12
	M9	87	19	58	93	52	71	59	47	11	43	20	26
Polished	M4	35	58	47	76	12	93	86	27	39	21	54	33
	M7	45	54	13	49	51	43	61	76	92	69	11	22
	M9	76	76	60	27	52	22	63	71	77	45	36	16
Check up	M5	37	90	49	58	62	17	91	27	9	51	42	44

B. Experimental Environment

All experiments were paper processor for Intel (R) Core (TM) i5-34700 CPU @ 3.20GHz, 4.00GB RAM, Windows 7 systems under the.

C. Computational results

The key problem with this algorithm is that selecting the appropriate parameters. As can be seen from the equation (9),the transition probability of ants value increases as α becomes larger. However, the value of α is too premature convergence Assembly prompted the

search trapped in local optimum; β values play an important role in the early running algorithms to process the shortest time. In under the condition of satisfying the constraint conditions, the goal of this article is to require the whole machining workshop of the shortest time and lowest cost and highest product quality. Figure 1 are the two parameters α and β on the influence of the characteristic curve. Obtained from the experiment, when the $\alpha = 0.4$, $\beta = 0.7$, we can obtain stable convergence of the Pareto front.

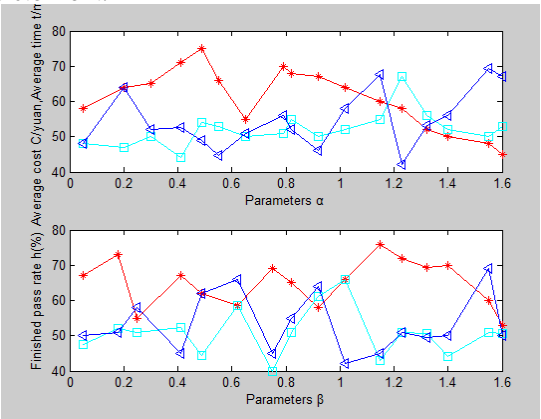


Figure 1. α and β algorithm characteristic curves

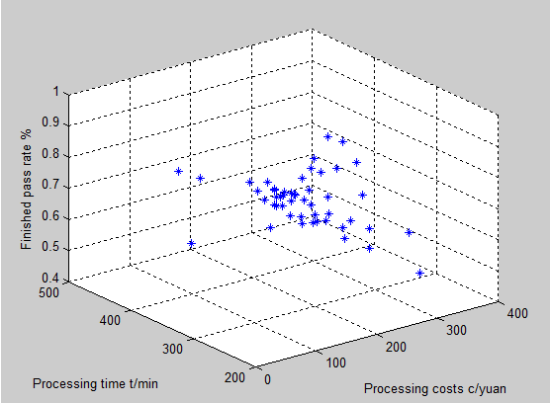


Figure 2. Flexible Job Shop Scheduling multi-objective optimization Pareto frontier

In summary, the constants initialized to $\alpha = 0.4$, $\beta = 7$, $\rho = 0.1$, $t = 300$, $\eta_{ij} = 0.5$. Below is using Matlab2011b multi-objective flexible job shop scheduling optimization simulation to obtain the Pareto frontier. Combined with lifting equipment processing characteristics of the coating workshop, according to the test data in the half moon, through the experiment we finally get the comprehensive optimum scheme of the shop scheduling.

In order to verify the algorithm in this paper to deal with the problem of multi-objective flexible shop scheduling optimization of superiority, at the same time, we used weighted coefficient under the condition of the same parameter Settings calculation comparison. Table 3 is two kinds of algorithm of operation 50 times the comparison result. Results show that compared with the weighted coefficient method, this algorithm in the processing time, processing solution on cost performance is good, can better solve the problem of multi-objective flexible shop scheduling.

TABLE III. COMPARISON OF TWO METHODS TABLE

Optimization goal	Herein algorithm		Weighting coefficient	
	Optimal solution	The average solution	Optimal solution	The average solution
Processing costs	*97*	*257*	*109*	*302*
Processing time	*207*	*310*	*221*	*339*
Finished pass rate	*90*	*75*	*86*	*73*

CONCLUSION

In this paper, we construct a flexible shop scheduling multi-objective optimization model based on the improved ant colony algorithm. The Processing cost, processing time, finished a pass rate as our optimization goals. We have improved the way pheromone update, so the algorithm while maintaining the diversity of Pareto sets, and can be a good approximation of the Pareto front. Simulation results show that the algorithm can effectively solve the multi-objective optimization problem. But how a more reasonable set of parameters in the ant colony algorithm to converge faster and worthy of further study.

REFERENCES

[1] Zhu Xinghui, Zhu Jinfu, Jiang Tao compare job shop scheduling problem several models of [J].Statistics and Decision, 2007, (23):174-176 DOI:10.3969 / j.issn.1002-6487.2007.23.066].

[2] HE Ting, COOKING Hai workshop production scheduling problem [J] Journal of Mechanical Engineering, 2000, 36 (5):97-102 DOI:10.3321 / j.issn:0577-6686.2000.05.025.

[3] Zhangchao Yong,DongXing, WangXiaojuan, etc.Improved Non - dominated Sorting Genetic Algorithm for Multi-objective flexible job shop scheduling [J] Journal of Mechanical Engineering, 2010, 46 (11):156-164 DOI:10.3901 / JME.2010.11.156

[4] P. Brucker and R. Schile, Job-shop scheduling with multipurpose machines, Computing, 45 (4) (1990) 369-375.

[5] A.Motaghedi, K. Sabri-Laghare and M. Heydari, solving flexible job shop scheduling problem with multi objectives, International Journal of Industrial Engineering and Production Research, 21 (2010)197-209.

[6] Q. Zhang, H. Manier and A. Manier, A genetic algorithm with tabu search procedure for flexible job shop scheduling with transportation constants and bounded proceding times, Computers and operations Research, 39 (2012) 1713-1723.

[7] Li JQ, Pan QK, Gao KZ (2011) Pareto-based discrete artificial bee colony algorithm for multi-objective flexible job shop scheduling problems. Int J Adv Manu Tech 55(9−12):1159−1169

[8] Zhang deposit, Zheng Pi-e, Xiao-Dan Wu colony and particle swarm optimization algorithm-based solution to multi-objective flexible scheduling problems [J] Journal of Computer Applications, 2007, 27 (4): 936-938

[9] Jiangjun Li, Pan Feng improved ant colony algorithm for solving multi-objective optimization problem [J] Southern Yangtze University: Natural Science, 2013, 12 (4): 394-398 DOI:10.3969 / j.issn.1671-7147.2013.04.004

[10] Shi marching focus of army, Tao delivery under penalty Pareto multi-objective flexible job shop scheduling optimization [J]. Journal of Mechanical Engineering, 2012, 48 (12):184-192 DOI:10.3901 / JME.2012.12.184.

An Improved QPSO Algorithm Based on Entire Search History

Ji Zhao[1,2]
1 Research Centre of Environment Science & Engineering
2 School of IoT Engineering
Jiangnan University
Wuxi, China
e-mail:queenji97@sohu.com

Yi Fu
School of IoT Engineering
Jiangnan University
Wuxi, China
e-mail:fy_bright@yahoo.com

Juan Mei
Research Centre of Environment Science & Engineering
Wuxi, China
e-mail:meijuanwx@gmail.com

***Abstract*—An improved QPSO algorithm based on entire search history (ESH-QPSO) is proposed. ESH-QPSO is an integration of the entire search history scheme and a standard quantum-behaved particle swarm optimization (QPSO).It guarantees that all updated positions are not re-visited before, which helps prevent premature convergence. The entire search history scheme partitions the continuous search space into sub-regions by using BSP tree. The partitioned sub-region servers as mutation range such that the corresponding mutation is adaptive and parameter-less. When sub-regions are formulated as which certain overlap exists between adjacent sub-regions, this allows particle move from a sub-region to another with better fitness. Compared with other traditional algorithms, the experiment results on 8 standard testing functions show that the proposed algorithm is superior regarding the optimization of multimodal and unimodal functions, with enhancement in both convergence speed and precision those demonstrate the effectiveness of the algorithm.**

Keywords: quantum-behaved particle swarm optimization; entire search history; adaptive mutate; binary space partitioning

I、 INTRODUCTION

In recently years, quantum-behaved particle swarm optimization algorithm (QPSO) [1]has attracted increasing attention because of its simple execution and superior performance. Similar to other global optimization algorithms[2], QPSO suffers from premature convergence and stagnate at local optimum. This is also a research focus in the swarm intelligence optimization algorithms and a lot of literatures studied on the issue [3][4]. In order to overcome the premature convergence of QPSO, it is important to increase the diversity of particles.

Several search algorithms [5-7] employ search history in the form of memory to adaptively guide the search strategies. However, they only use partial search histories – that is, only part of the information gained from the search is retained and the rest are discarded. Search history, including the performed operations, the positions of the evaluated solutions and the fitness values of the solutions, are valuable information to enhance the performance of a swarm intelligent algorithm (SI). Intuitively, it can be used to maintain diversity. It can also guide the search direction or suggest promising search regions of interest. In addition, when the same optimum reappears in the search history, it can warn that the search may have trapped in a local optimum. The non-revisiting genetic algorithm (cNrGA) is proposed by Yuen and Chow [8]. It is originally applied to genetic algorithm (GA) to prevent from solution re-evaluation. Meanwhile, the scheme also acts as a parameter-less mutation operation. The cNrGA is found to be more robust than GA. However, the adaptive mutation of cNrGA could not make gene fully explore due to sub-regions limitations. Moreover, because of the complexity of the GA algorithm, cNrGA has complex computation and slow convergence speed. Thus an improved QPSO algorithm based on entire search history is proposed.

This paper is organized as follows: Section II reports the details of QPSO-ESH. Section III presents the simulation setup. Section IV reports the simulation results. A conclusion is drawn in section V.

II、 QPSO ALGORITHM BASED ON ENTIRE SEARCH HISTORY

A. Entire search history scheme

Definition 1: Revisits Suppose E is a set of evaluated solutions, the solution x is a revisit if $x \in E$.

Definition 2: The Sub-region of solution x Suppose x is a solution in the search space S, i.e. $x \in S$, and S is partitioned into sub-region set $R = \cup_i r_i$ by a binary space portioning(BSP) tree T, the sub-region$r \in R$ is defined as the 'the sub-region of x' if $x \in r$ and r is represented by a leaf node of T.

The entire search history recorder scheme stores all visited solutions $\{e_i\}$ by a tree-structure archive, namely BSP tree. During iterations, the search space is being partitioned into set of regions R. In the BSP tree, a node represents a region of search space. Suppose a parent node P has two child nodes a and$b$. The child nodes linearly partition the sub-region of P into two overlapped sub-regions. The corresponding partitioning cuts along the k^{th} dimension where $k = \arg max|m(k) - n(k)|$. In this way, each previous solution generated by the QPSO is recorded in a node of the tree, and the BSP tree serves as an efficient data structure to query whether a new solution z is a revisit. In the whole solution process the BSP tree recordsall solutions in the search space and

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.26

analyzes the current solution by evolutionary search history of prior solutions. Since the tree construction depends on the sequence of solution set, the BSP tree is a random tree and its topology is different from trial to trial.

The scheme is analogous to a black box function (fig.1). The input x of that function can be any point in the search space. If x is a revisit, that function outputs solution x' such that 1)$x' \neq x$; 2)$x', x \in r_x \in R$. Otherwise, x' is assigned as x. Since the size of gradually decreases along with the iterations, and x' is randomly mutated from $r_x (r_x \in R)$, the expected distance between x' and x becomes smaller.

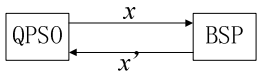


Fig.1 The scheme based on entire search history

B. An adaptive mutation mechanism

The node in the BSP tree linearly partition the adjacent sub-regions have certain overlap to each other, namely overlapped sub-regions. The search for the overlapped sub-region h of a solution x is implemented as a tree node search. The search starts from examining the root node whilst h is initialized as the whole search space. Each time the search moves downwards, h is contracted along a specified direction until the search reaches leaf node. Fig.2 summarizes the procedure to obtain the overlapped sub-region of a particle x. The sub-region contraction scheme guarantees that the resultant sub-region overlaps with all its adjacent sub-regions. Thus, the idea of overlapped sub-region together with mutation allows particles to gradually approach its optima nearby. If the solution x is a revisit and its sub-region is obtained and it is checked to be revisit, x will be replaced by the mutant of itself using One-Particle-Flip(OPF) mutation on that sub-region. This method is an extension variation of one-bit-flip mutation in common genetic algorithm (GA). Similar to one-bit-flip mutation, OPF mutates only one particle bit in the particle within the partition and this bit is randomly selected. OGF is a parameter-less adaptive mutation of ESH-QPSO, in the sense that the mutation is done randomly within the bounds of the particle bit, which are in turn defined by the partition. Suppose x is a N-dimensional particle to be mutated and $\prod_{i=1}^{N}[L(i), U(i)]$ is the mutation region of x. OPF starts from randomly selecting a dimension$j \in \{1,2,\ldots,N\}$. Then x is mutated as x' by replacing the j^{th} element of x with a random number in the range [L(j),U(j)]. The values of the particle bits in the rest of the dimensions are unchanged. The procedure of OPF is shown in Fig.3.

C. QPSO algorithm based on entire search history

Since the search space of ESH-QPSO is continuous, the number of possible solutions in a space partition, either for a small partition or a large partition, is infinite. No partition can be fully evaluated. Moreover the evolution mechanism is different between QPSO and GA, it is almost impossible to find the exactly same revisit in the sub-region. So a threshold e is set in ESH-QPSO.If the distance between the offspring solution and the pervious solution recorded in BSP tree is less than the threshold e, the solution is seen as an approximate solution and mutates adaptively.

Input: 1) BSP Tree T, 2) solution $x \in \Re^D$（D is dimension) and 3)search space S
$h = \Pi_j[l_j, u_j] := S$
Curr_node := root node of T
While (Curr_node has two child nodes m 和 n)
i=argmax$|m(i)-n(i)|$, i =1, 2, …, D
If ($|m(i)-x(i)| \leq |n(i)-x(i)|$)
Curr_node := child node m
$u_i := n(i)$
Else
Curr_node := child node n
$l_j := m(j)$
End
Loop
Output: the sub-region h of solution x

Fig.2 The pseudo-code of overlapping sub-regions for solutions x

Input: 1) the particle x need to mutate, 2)the mutation range $\prod_{i=1}^{N}[L(i), U(i)]$ of x
$x' := x$
Randomly selected dimension j $\{1,2,\ldots,D\}$.
$x'(j)$:= Rand([L(j),U(j)])
Output: x'

Fig.3The pseudo-code of OPF mutate in ESH-QPSO

Input: 1) a D-dimension minimization problem$F$(.)2) search space $S \subset \Re^D$3)the population size n
1. Initialize the current particle swarm $P=\{p_1,p_2,\ldots,p_n\}$
2. Initialize BSP tree T to which consists of root node only
3. Evaluate p_i: $f_i = F(p_i)$
4. Record $\{p_i\}$ to T
5. **While** terminate criteria is not satisfied
6. **For** i= 1,2,…,n
7. Update y_i by QPSO[1]
8. **If** the distance between y_i and recorded solution <$e$then
9.Search the overlopped sub-region of y_i,h_i
10.$\mathbf{y_i}$:= the mutant of $\mathbf{y_i}$ under OPF mutation
11. **EndIf**
12. **Next** i
13. ***gbest*** $=\boldsymbol{p_j} \in \boldsymbol{P}$where$j$ = arg min{ $F(p_i)$ }
14. **Loop**
Output：the optimal solution ***gbest***

Fig.4 The pseudo-code of ESH-QPSO

Fig.4 summarizes the procedure of ESH-QPSO. Given a D dimensional minimization problem F(.) with search space R^D, the algorithm starts from initializing the current population of n particles $P=\{p_1,p_2,\ldots,p_n\}$. Meanwhile, the BSP tree T is initialized to consist of the

root node. Then the population is then evaluated and is recorded by *T*. Afterwards, each particle in *P* is updated by QPSO and generated offspring particle population $\{y_1,y_2,\ldots, y_n\}$.Then , the BSP tree is accessed to check whether y_i is an approximate revisit. If y_i is an approximate revisit, its overlapped sub-region is obtained, and y_i is replaced by the mutant of itself using OPF mutation on that sub-region. The new particles population needs to be recalculated the particle's fitness value and inserted into the tree *T*. The processes are repeated until the termination criterion is satisfied.

III、 SIMULATION SETUP

A. Test Function

A set of standard test functions $F=\{f_1(x), f_2(x),\ldots, f_8(x)\}$ is adopted to test for the proposed algorithm and verify its performance.The eight test functions are shown in table 1.

B. Algorithms Setting

The performance of proposed ESH-QPSO is compared with standard QPSO, cNRGA and CLQPSO [9]. To ESH-QPSO, SQPSO and CLQPSO, the contraction expansion factor β decreases linearly from 1 to 0.5. To cNrGA, the cross rate is uniform crossover and sets as r_x=0.5; CLQPSO is set as [9]. The population size of all the algorithms is set to 100 and the maximum iteration number is set to 1000. The solution threshold *e* is shown in table 2. The dimension of all the test functions is set as D = 30. Each test function independently runs 30 times, and the mean optimum and standard deviations of test functions are obtained after 30 times experiments.

IV、 SIMULATION RESULTS

The performance of ESH-QPSO is compared with those of CLQPSO，SQPSO and cNrGA. Table 3 presents the means and standard deviations of the 30 runs of the four algorithms on the eight test functions. Seen from Table 3, ESH-QPSO is the most superior to other three algorithms for almost test functions exceptf_5 which get best perform by CLQPSO. The standard deviations of the optimal fitness for 30 independent trials of the algorithms are also listed in the table 3. The value in boldface indicates that the corresponding algorithm is the best amongst the test algorithms on a particular test function. It shows the standard deviation of ESH-QPSO is the best for all the test functions except for f_5This illustrates that for the vast majority of test functions, the stability of ESH-QPSO algorithm is the best. Therefore, from the detailed simulation results (mean and standard deviation), ESH-QPSO ranks first in five out of a total of 8 cases. Except for f_5 ESH-QPSO obtained the highest average accuracy and it also is the most stable.

V、CONCLUSION

Premature convergence and diversity are the most need to solve two problems by the QPSO algorithm. Therefor an improved QPSO algorithm based on entire search history (ESH-QPSO) is proposed that integrate the entire search history scheme and a standard quantum-behaved particle swarm optimization (QPSO). The two-dimensional space partitioning tree (BSP) is used to record the solutions in the process of evolution, which ensures that each update particle position will not be revisited and prevents the algorithm fall into the local convergence and prevents premature. The continuous search space is divided into different sub-regions as a particle mutation range by the BSP tree. This makes the corresponding mutation is a kind of adaptive mutation and provides guidance for .global and local search of particles.The overlap between adjacent sub-regions also provides a channel for particles move in the adjacent area result in the particles can be more easily move from the area to a better area. Compared with other traditional algorithms, the experiment results on 8 standard testing functions show that the proposed algorithm has the best optimization ability, with enhancement in both convergence speed and precision those demonstrate the effectiveness of the ESH-QPSO.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from “Qing LanProject[2012-16]” of Jiangsu province and National Natural Science Foundation of China 61300149

REFERENCE

[1] Sun J, Fang W, Wu X J, et al. Quantum-behaved particle swarm optimization: Theory and application[M]. Beijing: Tsinghua University Press, 2011: 49-50

[2] Kennedy J, Eberhart R C.Particle swarm optimization[C]. Proc of IEEE IntConf on Neural Networks. Perth: IEEE Press, 1995, pp: 1942-1948.

[3] WU Tao;YAN Yu-song;CHEN Xi; Improved dual-group interaction QPSO algorithm based on random Evaluation[J]. Control and Decision, 2015, 30(3):526-530

[4] Sun, J, Fang, W, Wu, X, Palade, V, Xu, W.Quantum-Behaved Particle Swarm Optimization: Analysis of Individual Particle Behavior and Parameter Selection[J].Evolutionary Computation, 2012, 20(3):349-393

[5] Glover, M.Laguna, TabuSearch[M], Kluwer Academic Publishers, 1997.

[6] R.G.Reynolds.An overview of cultural algorithms in, McGraw Hill Press, [J].Advances in Evolutionary Computation, 1999

[7] J.D.Farmer, N.Packard and A.Perelson. The immune system, adaptation and machine learning[J].Physical D, 1986, 2:187-204

[8] S.Y. Yuen and C. K. Chow, “Continuous non-revisiting genetic algorithm[C]” in Proc. IEEE Congr. Evol. Comput., 2009, pp: 1896-1903.

[9] W. Chen,D. Zhou, J. Sun,W.B. Xu;Improved quantum-behaved particle swarm optimization algorithm based on comprehensive learning strategy[J], Control and Decision, 2012, 27(5):719-723.

Table 1 Test Functions

		X	x_0	y_0
f_1Spherical model	$\sum_{i=1}^{D} x_i^2$	[-100，100]^D	[0,…,0]	0
f_2 Schwefel's Problem 2.22	$\sum_{i=1}^{D} \lvert x_i \rvert + \prod_{i>1}^{D} \lvert x_i \rvert$	[-10，10]^D	[0,…,0]	0
f_3 Schwefel's Problem 1.2	$\sum_{i=1}^{D} \left\{ \sum_{j=1}^{i} x_j \right\}^2$	[-100，100]^D	[0,…,0]	0
f_4 Schwefel's Problem 2.21	$\max_{i\in[1,D]} \lvert x_i \rvert$	[-10，10]^D	[0,…,0]	0
f_5 Restrigin's function	$\sum_{i>1}^{D} [x_i^2 - 10\cos(2\pi x_i) + 10]$	[-5.12，5.12]^D	[0,…,0]	0
f_6 Ackley	$-20exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right) - exp\left(\frac{1}{D}\sum_{i=1}^{D} cos2\pi x_i\right)$+20+$e$	[-32，32]^D	[0,…,0]	0
f_7Rosenbrock'sfuncton	$\sum_{i=1}^{D-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2$	[-29，31]^D	[0,…,0]	0
f_8Griewank function	$\frac{1}{4000}\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} cos\frac{x_i}{\sqrt{i}} + 1$	[-600，600]^D	[0,…,0]	0

Table 2 The solutions threshold *e* settings

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
threshold *e*	1e-7	0	5e-2	5e-2	1e-1	1e-3	5e-3	1e-3

Table 3The optimize results of the test function

	f_1	f_2	f_3	f_4
SQPSO	1.47998E-14	1.11684E-10	186.3872131	0.510266323
	(3.04472E-14)	(4.15217E-10)	(104.0295677)	(0.222604474)
ESH-QPSO	**3.95622E-17**	**6.3487E-13**	**179.3536971**	**0.178691122**
	(8.46183E-17)	**(2.35457E-12)**	**(75.41568806)**	**(0.090119977)**
cNrGA	0.091400318	0.014986892	3353.915859	2.374221208
	(0.275677938)	(0.012854059)	(1894.914266)	(2.314962811)
CLQPSO	0.24313441	0.047356203	17341.37742	34.9262655
	(0.080794574)	(0.008233747)	(2989.341129)	(2.014683633)
	f_5	f_6	f_7	f_8
SQPSO	15.71033071	1.95029E-08	25.71260941	0.008289674
	(4.179950722)	(2.28787E-08)	(1.573154449)	(0.009056309)
ESH-QPSO	17.26429913	**8.97891E-09**	**24.16385639**	**0.005907779**
	(4.02561364)	**(8.66625E-09)**	**(1.538993104)**	**(0.008410973)**
cNrGA	0.845910299	0.069056062	35.22865152	0.11359066
	(1.20528348)	(0.240890393)	(20.92783323)	(0.145049358)
CLQPSO	**0.261776014**	0.606512754	248.6310265	0.423170878
	(0.094784807)	(0.243677983)	(42.63673184)	(0.078499638)

Proportional Fairness based Resources Allocation Algorithm for LEO Satellite Networks

Shuang Xu, Xingwei Wang, Min Huang
College of Information Science and Engineering,
Northeastern University
Shenyang, China
xiaoshuang_0320@163.com; wangxw@mail.neu.edu.cn; mhuang@mail.neu.edu.cn

***Abstract*—Low Earth Orbit (LEO) satellite networks play an important role in global real-time satellite communication. However, due to the limitations of satellite bandwidth and power, it is urgent to utilize resources efficiently and prevent unnecessary resources waste. Existing researches usually focus on resources allocation for downlinks without considering Inter-Satellite Links (ISLs). This paper introduces proportional fairness by combining ground stations number with user data links life time and formulates the resources allocation problem among ISLs as a nonlinear mixed integer programing problem. Then Proportional Fairness based Resources Allocation algorithm is proposed to obtain the optimal solution using Swallow Swarm Optimization. It achieves proportional fairness among ISLs by sacrificing node throughput capacity. Moreover, the impact of counter-rotating seam and latitude regions is analyzed.**

Keywords-proportional fairness; LEO satellite networks; resources allocation; swallow swarm optimization

I. INTRODUCTION

With the development of global real-time service, smaller antennas, and lower power supplied by smart terminals, Low Earth Orbit (LEO) satellite networks with Inter-Satellite Links (ISLs) which provides short latency and less dependence on terrestrial networks, should be a better choice [1]. However, its satellite power and bandwidth resources are scarce. Furthermore, its mobility and the large number of satellites required for global coverage bring great challenges for resources management. As a result, it is crucial to devise an effective resources allocation algorithm for LEO satellite networks to improve the resources utilization and meanwhile maximize the network capacity.

A dynamic bandwidth allocation scheme based on traffic distributions and channel conditions for the downlinks of multi-spot-beam satellite system has been proposed [2], in which a trade-off between the maximum total capacity and fairness among the spot beams is considered. It achieves the proportional fairness resulting in loss of total system capacity. To improve total system capacity, [3] gives the priority to active beam to compensate the degradation of total system capacity. Based on QoS requirements and delay of LEO small satellite system, the utility-base sub-carrier power allocation strategy for downlinks has been proposed in [4]. It balances efficiency and fairness and satisfies the users QoS requirements. [2-4] just optimized the allocation of power or bandwidth resulting in a waste of the other. In [5] the optimal power and bandwidth allocation is researched considering the delay of real-time traffic of satellite network downlinks. It achieves trade-off between throughput and delay. In [6] the joint bandwidth and power allocation algorithm is proposed which improves the total system capacity and fairness among spot beam. Although, [5, 6] allocate bandwidth and power jointly, they mainly solve the resource allocation for downlinks without considering the ISLs.

This paper proposes a Proportional Fairness based Resources Allocation (PFRA) algorithm for LEO satellite networks. It not only jointly allocates bandwidth and power considering ISLs, but also combines ground stations number and User Data Links (UDLs) life time to improve the fairness. Simulation results demonstrate it achieves proportional fairness among ISLs at a cost of node throughput capacity. Moreover, we discuss the impact of counter-rotating seam and latitude regions on node throughput capacity.

II. PROPORTIONAL FAIRNESS BASED BANDWIDTH AND POWER ALLOCATION ALGORITHM

A. The Network Scenario

Each satellite interconnects with its adjacent satellites by ISLs in LEO satellite networks. Due to the mobility of LEO satellites, the coverage region, relay service requirements and UDLs life time of each LEO satellite are different and time varying. The ground stations number in the footprint of each satellite is affected by the traffic demands and users number, so the ground stations are non-uniform distribution.

These result in three intuitive constraints: (1) The more relay service requirements, the more bandwidth and power should be allocated to the corresponding relay LEO satellites to provide sufficient links capacity; vice versa. (2) The longer UDLs lifetime, the longer time they would be available to transmit data. More bandwidth and power should be allocated to the corresponding satellite to reduce the cost of establishing UDLs; vice versa. (3) The more ground stations in the satellite footprint, the more data need to be transmitted. Thus, more bandwidth and power should be allocated to the corresponding satellite; vice versa.

B. The User Data Link Lifetime Constraint

Ground stations move in and out of satellite coverage with satellites moving. UDL lifetime $T_{lifespan}$ refers to the time

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.27

period that from the current time t_{now} when ground station G_0 is in the coverage of satellite S_0 to the time t_{end} when G_0 is out of the coverage of S_0. The maximal UDL lifetime $T_{maxLifespan}$ refers to the time period that from time t_{begin} when the UDL can be established to the time t_{end}. We have

$$T_{lifespan}=t_{end}-t_{now} \quad (1)$$

$$T_{maxLifespan}=t_{end}-t_{begin} \quad (2)$$

C. Optimization Problem Formulation

For satellite $s_{i_0 j_0}$, suppose its total bandwidth is B which is divided into $N_{subchan}$ orthogonal sub-channels. Define the satellites that establish ISLs with $s_{i_0 j_0}$ as its adjacent satellites, and adjacent satellites number is N_{nei}. The bandwidth allocation is transformed into sub-channels allocation. The sub-channels allocation result can be denoted by matrix C:

$$C=\begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,N_{subchan}} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,N_{subchan}} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N_{nei},1} & c_{N_{nei},2} & \cdots & c_{N_{nei},N_{subchan}} \end{bmatrix}$$

where $c_{ki} \in \{0,1\}$, $c_{ki}=1$ denotes $i\text{-}th$ sub-channel is allocated to $k\text{-}th$ adjacent satellite, and all the sub-channels allocated to $k\text{-}th$ adjacent satellite is denoted by set Ω_k. The power allocation result is denoted by $P=[p_1,p_2,\cdots,p_{N_{subchan}}]$, here, p_i represents the transmitter power allocated to $i\text{-}th$ sub-channel.

We define the ISLs capacity T_k^{link} between satellite $s_{i_0 j_0}$ and $k\text{-}th$ adjacent satellite as the total capacity of sub-channels that is allocated to $k\text{-}th$ adjacent satellite. Thus

$$T_k^{link}=\frac{B}{N_{subchan}}\sum_{i=1}^{N_{subchan}} c_{ki}\log_2(1+SNR_{k,i}) \quad (3)$$

where $SNR_{k,i}$ represents SNR of $i\text{-}th$ sub-channel of $k\text{-}th$ adjacent satellite.

We define node throughput capacity $NTC_{i_0 j_0}$ as the total ISLs capacity between $s_{i_0 j_0}$ and its all adjacent satellites.

$$NTC_{i_0 j_0}=\sum_{k=1}^{N_{neig}} T_k^{link}=\frac{B}{N_{subchan}}\sum_{k=1}^{N_{nei}}\sum_{i=1}^{N_{subchan}} c_{ki}\log_2(1+SNR_{k,i}) \quad (4)$$

Based on our three intuitive constraints we introduce proportional constraint by combining ground stations number with UDLs life time. We formulate the problem as a nonlinear mixed integer programing problem.

$$\begin{aligned} &\max\ NTC_{i_0 j_0} \\ &\sum_{k=1}^{N_{nei}} c_{ki}=1 \quad i=1,\cdots,N_{subchan} \\ &c_{ki}\in\{0,1\} \quad k=1,\cdots,N_{nei}; i=1,\cdots,N_{subchan} \\ &\sum_{i=1}^{N_{subchan}} p_i \le p_{total} \\ &p_i\in[0,p_{total}] \quad i=1,\cdots,N_{subchan} \\ &SNR_i \ge SNR_{min} \quad i=1,\cdots,N_{subchan} \\ &T_k^{link}:T_l^{link}=\varphi_k:\varphi_l,\forall k,l\in\{1,\cdots,N_{nei}\},k\ne l \end{aligned} \quad (5)$$

where

$$\varphi_k=\alpha_{relay}+\beta_{groundStation}\cdot\sum_{i\in\Omega_G}\frac{T_{lifespan}^{ki}}{T_{\max Lifespan}},k=1,2,\cdots,N_{neighbor} \quad (6)$$

where α_{relay} and $\beta_{groundStation}$ represent influence factor of relay service and ground stations numbers, respectively.

D. Heuristic Approach to Optimum Bandwidth and Power

Compared with particle swarm optimization and fish swarm optimization, SSO [7] has high efficiency. Thus, it is used to solve the optimization problem.

1) Dynamic Feasible Region

Feasible solutions are generally considered better than the infeasible. While at later evolution there are basically feasible solutions with poor quality. The infeasible solutions with less constraint violation and better objective function value could provide more valuable information for searching optimum solution than the feasible with poor quality. Thus, we introduce dynamic feasible region that is composed of feasible region and infeasible region to avoid trapping in local optimum. Structuring a dynamic feasible region need to make rules of infeasible solutions size change and quality evaluation. At earlier evolution, the infeasible solutions size N_{select} is enlarged to enhance population diversity. While the optimum solution must be feasible solution, hence N_{select} should be narrowed with the iteration process.

$$N_{select}=N_{leader}\cdot(1-\frac{iter}{maxIter}) \quad (7)$$

where N_{leader} is leader swallows number, *iter* is current iteration, *maxIter* is maximum iteration.

The quality evaluation of infeasible solutions depends on objective function $Obj(X)$ and constraint violation $Con(X)$. We turn the constraints into inequality and equality.

$$g_1(X):=(\sum_{i=1}^{N_{subchan}} p_i-p_{total})+\sum_{i=1}^{N_{subchan}}(SNR_{\min}-SNR_i)\le 0 \quad (8)$$

$$g_2(X):=\sum_{i=1}^{N_{subchan}}(\sum_{k=1}^{N_{nei}} c_{ki}-1)+\sum_{k\ne l}^{N_{nei}}(\frac{T_k^{link}}{T_l^{link}}-\frac{\varphi_k}{\varphi_l})=0 \quad (9)$$

and we define $Obj(X)$ and $Con(X)$ as:

$$Obj(X)=NTC_{i_0 j_0} \quad (10)$$

$$Con(X)=(g_1(X))^2+(g_2(X))^2 \quad (11)$$

Given that the values of $Obj(X)$ and $Con(X)$ may be in different order of magnitude, we normalize them as follows:

$$F(X)=\frac{Obj(X)-Obj_{min}}{Obj_{max}-Obj_{min}} \quad (12)$$

$$G(X)=\frac{Con_{max}-Con(X)}{Con_{max}} \quad (13)$$

$Obj_{max}=\max_{i=1,\cdots,N_{population}}(Obj(X_i))$, $Obj_{min}=\min_{i=1,\cdots,N_{population}}(Obj(X_i))$, $Con_{max}=\max_{i=1,\cdots,N_{population}}(Con(X_i))$, $N_{population}$ is population size.

We defined the infeasible solutions evaluation function,

$$Estimate(X_{infeasible})=F(X_{infeasible})\cdot(G(X_{infeasible}))^{(0.5+iter/maxIter)} \quad (14)$$

Sort the infeasible solutions according to their evaluation function values, then select N_{select} infeasible solutions at the top of queue to compose dynamic feasible region.

2) Fitness Function and Initialization

Regard infeasible solutions in $D_{feasible}$ as feasible solutions and neglect their constraint violations, the fitness function

$$Fit(X)=\begin{cases} F(X) & X \in D_{feasible} \\ G(X)-1 & X \notin D_{feasible} \end{cases} \quad (15)$$

In SSO, swallows correspond to problem solutions. The $N_{population}$ swallows are generated randomly as initial solutions. Each swallow $X=(\vec{c},\vec{p})$ is composed of binary vector $\vec{c}=(c_{1,1},\cdots,c_{1,N_{subcahn}},\cdots,c_{N_{nei},1},\cdots,c_{N_{nei},N_{subchan}})$ and real vector $\vec{p}=(p_1,\cdots,p_{N_{subchan}})$. Similarly the velocity of each swallow is also generated randomly.

3) Leader Swallow

Divide swallow population into some subpopulations. Select N_{leader} better swallows as leader swallows according to their fitness function values. The leader with best fitness function value is named as Head Leader (HL). It is the major leader in the entire population and guides all the swallows to the public optimum point. The others are Local Leaders (LL) which conduct the internal subpopulations and guide the other members to local optimum points. In every iteration, leader swallows are changed according to new swallows sort.

4) Explorer Swallow

Explorer swallows are responsible for exploring in problem space and make random move regarding V_{LL}^i (the velocity vector of swallow toward HL) and V_{LL}^i (the velocity vector of swallow toward LL).

Denote the real vector and binary vector of V_{HL}^i by V_{PHL}^i and V_{CHL}^i, then for V_{PHL}^i

$$V_{PHL}^{i+1}=V_{PHL}^i+\alpha_{HL}\cdot rand()\cdot(X_p^{HistoryBest}-X_P^i)+\beta_{HL}\cdot rand()\cdot(X_P^{HL}-X_P^i) \quad (16)$$

$$\alpha_{HL}=(1-\frac{Fit(X^i)+1}{Fit(X^{HL})+1})\cdot[(a_1-a_2)\cdot\frac{iter}{maxIter}+a_2] \quad (17)$$

$$\beta_{HL}=(1-\frac{Fit(X^i)+1}{Fit(X^{HL})+1})\cdot[(a_2-a_1)\cdot\frac{iter}{maxIter}+a_1] \quad (18)$$

where X_P^i is real vector of swallow current position, $X_P^{HistoryBest}$ is real vector of the best position from beginning up to now, X_P^{HL} is real vector of HL current position, α_{HL} and β_{HL} are acceleration coefficients, and $a_1=0.5$, $a_2=2.5$.

For binary vector, we defined an operation *BinOper* :

Step 1: Convert input vectors $X^1,X^2,\cdots,X^n$ into $N_{nei}\times N_{subchan}$ matrices $X_1^{'},X_2^{'},\cdots,X_n^{'}$;

Step 2: Successively compare every column in $X_1^{'},X_2^{'},\cdots,X_n^{'}$, if *l-th* columns in $X_1^{'},X_2^{'},\cdots,X_n^{'}$ satisfy $X_{1l}^{'}=X_{2l}^{'}=\cdots=X_{nl}^{'}$, then set *l-th* column of X matrix $X_l=X_{1l}^{'}$; otherwise, calculate $X_{suml}=X_{1l}^{'}+X_{2l}^{'}+\cdots+X_{nl}^{'}$;

Step 3: Analyze elements of X_{suml}, if there is only one largest element which locates in *n-th* row, then set the element of X in *n-th* row be 1 and others be 0; otherwise, randomly set an element be 1 and the others be 0.

For V_{CHL}^i, $V_{CHL}^{i+1}=BinOper(V_{CHL}^i,X_C^{historyBest},X_C^{HL},X_C^i)$.

Similarly, denote the real vector and binary vector of V_{LL}^i by V_{PLL}^i and V_{CLL}^i, then for V_{PLL}^i

$$V_{PLL}^{i+1}=V_{PLL}^i+\alpha_{LL}\cdot rand()\cdot(X_p^{HistoryBest}-X_P^i)+\beta_{LL}\cdot rand()\cdot(X_P^{LL}-X_P^i) \quad (19)$$

$$\alpha_{LL}=(1-\frac{Fit(X^i)+1}{Fit(X^{LL})+1})\cdot[(a_1-a_2)\cdot\frac{iter}{maxIter}+a_2] \quad (20)$$

$$\beta_{LL}=(1-\frac{Fit(X^i)+1}{Fit(X^{LL})+1})\cdot[(a_2-a_1)\cdot\frac{iter}{maxIter}+a_1] \quad (21)$$

For V_{CLL}^i, $V_{CLL}^{i+1}=BinOper(V_{CLL}^i,X_C^{historyBest},X_C^{LL},X_C^i)$.

Update the real vector of V^{i+1} by $V_P^{i+1}=V_{PHL}^{i+1}+V_{PLL}^{i+1}$ and the binary vector of V^{i+1} by $V_C^{i+1}=BinOper(V_{CHL}^i,V_{CLL}^i)$. Update the real vector of X^{i+1} by $X_P^{i+1}=X_P^i+V_P^{i+1}$ and the binary vector of X^{i+1} by $X_C^{i+1}=BinOper(X_C^i,V_C^{i+1})$.

5) Aimless Swallow

Select swallows with relatively bad fitness function values as aimless swallows. Their duty is to explore neglected space. They move randomly as follows:

For the real vector of aimless swallow position,

$$X_P^{i+1}=X_P^i+[rand\{-1,1\}\cdot\frac{rand(0,p_{total})}{1+rand()}] \quad (22)$$

For the binary vector of aimless swallow position,

$$X_C^{i+1}=BinOper(X_C^i,C_{rand}) \quad (23)$$

where C_{rand} is sub-channels allocation vector generated randomly.

6) Flowchart of PFRA

The flow of PFRA is depicted in Fig. 1:

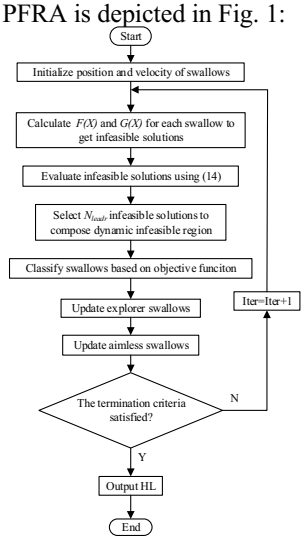


Figure 1. Flowchart of PFRA

III. SIMULATIONS AND PERFORMANCE EVALUATION

We evaluate the performance of PFRA using the Iridium [8] satellite constellation system on STK and Eclipse 4.2. Its satellite power and bandwidth are respectively 20W and 200 Mbps. We randomly deploy 100 ground stations on the earth

and uniformity set 154 sample points in a period. The average node throughput capacity is the average value of node throughput capacity of 154 sample points.

A. Impact of the Counter-rotating Seam

The satellites along the counter-rotating seam are termed boundary satellites and the others termed non-boundary satellites. Fig. 2 shows the node throughput capacity of boundary and non-boundary satellites (their coverage areas are similar) in a period using the proposed PFRA. It is shown that node throughput capacity of non-boundary satellite has a larger fluctuation than boundary satellite, but their curve variation tendencies are approximate and unanimous. The counter-rotating seam has a strong influence on fluctuation ranges and a weak influence on variation tendencies.

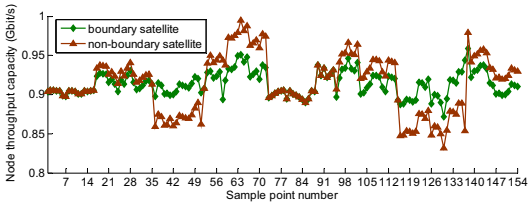


Figure 2. Node throughput capacity of boundary and non-boundary satellites

B. Performance Evaluation of the PFRA

We compare PFRA with MCRA which maximizes the node throughput capacity just under the limitations of total bandwidth and power to allocate bandwidth and power. As shown in Fig. 3, the curve of MCRA is stable while the satellite is in low altitude region. When the satellite gradually moves from low altitude region to mid latitude region, node throughput capacity increases rapidly. When the satellite moves into polar region, the inter-plane ISLs are broken so that the curve falls into low. As the satellite moves into the southern hemisphere, the curve shows similar variation tendencies. However, the curve of PFRA randomly fluctuates and when the satellite is in polar region the curve falls into the lowest ebb. The reason is that PFRA considers the influence of ground stations number and UDLs life time.

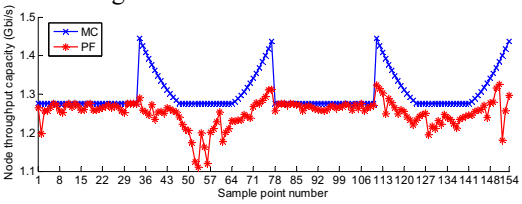


Figure 3. Comparison of node throughput capacity of MCRA and PFRA

Fig. 4 shows the average node throughput capacity of MCRA and PFRA under different power: 20W, 60W, 100W. As the power increases, the average node throughput capacity of MCRA and PFRA both increase. While their increase amplitudes decrease with the power increasing. MCRA achieves greater average throughput capacity than PFRA. While PFRA achieves proportional fairness among ISLs at a cost of node throughput capacity.

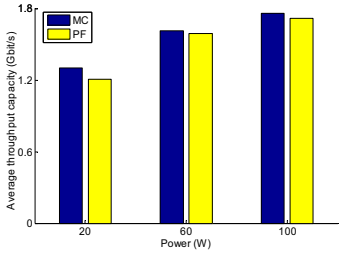


Figure 4. Average node throughput capacity under different power

IV. CONCLUSIONS

We formulate the problem as a nonlinear mixed integer programing problem considering the proportional fairness among ISLs and propose PFRA. Simulation results demonstrate that PFRA achieves proportional fairness sacrificing node throughput capacity. The impact of the counter-rotating seam on throughput capacity variation tendency can be neglected. Node throughput capacity fluctuates with the change of latitude regions.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation for Distinguished Young Scholars of China under Grant No. 61225012 and No. 71325002; the Specialized Research Fund of the Doctoral Program of Higher Education for the Priority Development Areas under Grant No. 20120042130003; Liaoning BaiQianWan Talents Program under Grant No. 2013921068.

REFERENCES

[1] F. Alagoz and G. Gur, "Energy efficiency and satellite networking: a holistic overview," Proceedings of the IEEE, vol. 99, Nov. 2011, pp. 1954-1979, doi: 10.1109/JPROC.2011.2165192.

[2] U. Park, H. W. Kim, D. S. Oh and B. J. Ku, "A dynamic bandwidth allocation scheme for a multi-spot-beam satellite system," Etri Journal, vol. 34, Aug. 2012, pp. 613-616, doi: 10.4218/etrij.12.0211.0437.

[3] U. Park, H. W. Kim, D. S. Oh and B. J. Ku, "Flexible bandwidth allocation scheme based on traffic demands and channel conditions for multi-beam satellite systems," 2012 IEEE Vehicular Technology Conference (VTC fall), IEEE Press, Sep. 2012, doi: BEJ90.

[4] W. Xiaolu, C. Yueyun, G. Liqiang and M. Chao, "A utility-based OFDM resource allocation scheme for LEO small satellite system," Cyberspace Technology on Cyberspace Technology 2013 (CCT 2013), IET Press, Nov. 2013, pp. 68-73, doi: 10.1049/cp.2013.2096.

[5] Z. Ji, Y. Z. Wang, W. Feng and J. H. Lu, "Delay-aware power and bandwidth allocation for multiuser satellite downlinks," IEEE Commun. Lett, vol. 18, Nov. 2014, pp. 1951-1954, doi: 10.1109/LCOMM.2014.2363111.

[6] H. Wang, A. Liu and X. Pan, "Optimization of joint power and bandwidth allocation in multi-spot-beam satellite communication systems," Mathematical Problems in Engineering, vol. 2014, 2014, pp. 1-9, doi: 10.1155/2014/683604.

[7] M. Neshat, G. Sepidnam and M. Sargolzaei, "Swallow swarm optimization algorithm: a new method to optimization," Neural Computing & Applications, vol. 23, Aug. 2013, pp. 429-454, doi: 10.1007/s00521-012-0939-9.

[8] LEO Communications Satellites: The IRIDIUM Constellation, http://ccar.colorado.edu/asen5050/projects/projects_2000/redlin/.

Quantum-behaved Particle Swarm Optimization with Cooperative Coevolution for Large Scale Optimization

Na Tian
Department of Educational Technology, Jiangnan University, Wuxi 214122, China

***Abstract*—Quantum-behaved particle swarm optimization (QPSO) has successfully been applied to unimodal and multimodal optimization problems. However, with the emerging and popular of big data and deep machine learning, QPSO encounters limitations with high dimensions. In this paper, QPSO with cooperative coevolution (QPSO_CC) is used to decompose the high dimensional problems into several lower dimensional problems and optimize them separately. The numerical experimental results show that QPSO_CC has comparative or even better performance than other algorithms.**

Keywords-large scale; quantum-behaved particle swarm optimization; cooperative coevolution; domain decompositimponent

I. INTRODUCTION

As we all know, performance of the stochastic optimization algorithms (including particle swarm optimization (PSO), genetic algorithms (GA)) deteriorates as the dimensionality of the search space increases. A natural approach to deal with high-dimensional optimization problems is to adopt a divide-and-conquer strategy. Clearly, the effectiveness of such approach depends heavily on the decomposition strategies used. Especially for nonseparable problems, because the interdependencies among different variables could not be captured well enough.

In [10], Potter firstly suggested that the search space should be partitioned into smaller vectors, and found that the decomposition lead to a significant improvement in performance over the classic GA. An attempt to apply Potter's CC model to PSO is made in [11], where two cooperative models, CPSO- S_k and CPSO- H_k were developed. Yang et al. [12] proposed a new decomposition method based on random grouping and adaptive weighting to deal with nonseparable problems. [13] reveals that it is even more beneficial to apply random grouping more frequently. The recent work in [14] proposed a new PSO cooperative coevoluation (CC) framework with ring topology (lbest), new strategy to update personal best and global best vector, and dynamically changing group size, which shows competitive performance and scale up to 2000 dimensions.

QPSO, proposed by Sun in 2004 [15], has been proven to perform better than PSO both in exploration and exploitation abilities [16-17]. However, improved variants of QPSO were just tested on low dimensional problems (10, 20, 30). The scalability to high dimensions have not been validated. Building on the previous works, this paper gives a cooperative coevolutionary framework on QPSO (QPSO_CC) to solve large scale optimization problems.

The rest of the paper is organized as follows. The details of the QPSO_CC are presented in Section 2. Section 3 describes the experimental setup, results and analysis. Finally section 4 gives the conclusion.

II. QPSO WITH COOPERATIVE COEVOLUTION

A minimization problem is assumed in this section:

$$\min f(x), x \in \Omega \tag{1}$$

where $\Omega \subset R$ is the search space, x is a vector with dimensional D.

A. QPSO

In quantum world, the velocity of the particle is meaningless, so in QPSO system, position is the only state to depict the particles, which moves according to the following equation[15]:

$$X_i(t+1) = p_i(t) \pm \alpha \left| mbest(t) - X_i(t) \right| \ln\left(1/u\right), \tag{2}$$

where u is a random number uniformly distributed in $(0,1)$, $mb(t)$ called Mean Best Position, is defined as the mean value of personal best positions of the swarm:

$$mb(t) = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1}(t), \frac{1}{M}\sum_{i=1}^{M} P_{i2}(t), ..., \frac{1}{M}\sum_{i=1}^{M} P_{iD}(t) \right) \tag{3}$$

The parameter α in Eq. (2) is named as Contraction-Expansion (CE) coefficient, which can be adjusted to control convergence rate. The most commonly used method to control CE is linearly decreasing from $\alpha_{\max}$ to $\alpha_{\min}$:

$$\alpha = (\alpha_{\max} - \alpha_{\min})(t_{\max} - t)/t_{\max} + \alpha_{\min} \tag{4}$$

where t is the current iteration step, $t_{\max}$ is the predefined maximum iteration steps, $\alpha_{\max}$ and $\alpha_{\min}$ are the maximum and minimum value of CE.

B. Cooperative coevolution framework

In CC, the search space is decomposed into smaller components and each of them is assigned to a subpopulation. The subpopulations are evolved mostly separately with the only cooperation during fitness evaluation. The general framework of CC is described as follows:

(1) Decompose a vector into K lower dimensional subcomponents.
(2) For $i = 1:K$
(3) Optimize the i th subcomponent with QPSO for a predefined number of iterations.
(4) End For

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.28

(5) Stop if halting criteria are satisfied; otherwise go to step (2).

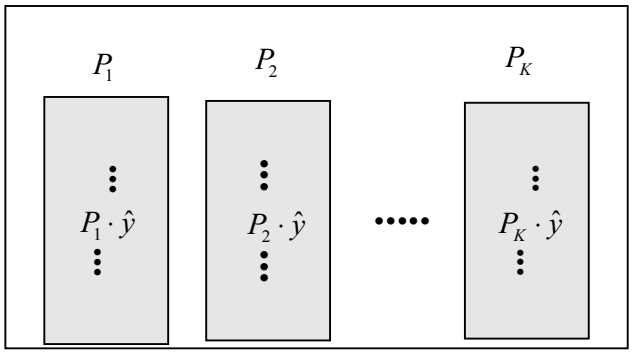


Figure 1 Concatenation of $P_1 \cdot \hat{y}, P_2 \cdot \hat{y}, ..., P_i \cdot \hat{y}, ..., P_K \cdot \hat{y}$ to constitute global best $\hat{y}$

C. context vector

In step (3), evaluation of the i th subcomponent cannot be computed directly. A context vector is required to provide a suitable context in which a subcomponent can be evaluated. The simplest scheme is to take the global best particles from each of the K swarms and concatenate then to form a D-dimensional vector (as shown in Fig. 1). To calculate the fitness of particles in swarm i, the other $D-1$ components in the context vector are kept unchanged, while the i th component of the context vector is replaced in turn by each particle from swarm i. The concatenation of the i th subcomponent with context vector is defined as:

$$b(i,z) \equiv \left(P_1 \cdot \hat{y}, P_2 \cdot \hat{y}, ..., P_{i-1} \cdot \hat{y}, z, P_{i+1} \cdot \hat{y}, ..., P_K \cdot \hat{y}\right) \quad (5)$$

in which, $P_i \cdot \hat{y}$ is the global best particle in swarm i, z represents the position of any particle in swarm P_i, $P_i \cdot x_j$ refers to the position of particle j in swarm i.

The idea is to evaluate how well $P_i \cdot x_j$ cooperates with the best individuals from all the other swarms.

D. Random decomposition strategy

To handle the high dimensional non-separable problems and put the correlated variables into the same subcomponent, a random decomposition strategy is used in QPSO_CC.

In the traditional CC frame, the D-dimensional search space is decomposed into K subcomponents, each corresponding to a swarm of s -dimensions (where $D = K * s$). Since we do not know in advance how these K subcomponents are correlated for any given problem, such static grouping method is likely to put some interacting variables into different subcomponents.

The random decomposition is the simplest dynamic grouping method and does not require any prior knowledge of the problems to be solved, in which, each subcomponent is constructed by randomly selecting s -dimensions from the D-dimensional search space.

2.3 Pseudo-code of QPSO_CC

repeat
Randomly permute all D dimension indices ;
Initialize K swarms, each with s dimensions randomly chosen from D, and $D = K * s$;
the i th swarm is denoted as $P_i, i \in [1,2,...,K]$;
for each swarm $i \in [1,2,...,K]$
  for each particle $j \in [1,...,swarmSize]$
    if $f\left(b\left(i, P_i \cdot x_j\right)\right) < f\left(b\left(i, P_i \cdot y_j\right)\right)$ then
      $P_i \cdot y_j \leftarrow P_i \cdot x_j$;
    if $f\left(b\left(i, P_i \cdot y_j\right)\right) < f\left(b\left(i, P_i \cdot \hat{y}\right)\right)$ then
      $P_i \cdot \hat{y} \leftarrow P_i \cdot y_j$;
  end for
end
for each swarm $i \in [1,2,...,K]$
  for each particle $j \in [1,...,swarmSize]$
    Perform position update for the j th particle in swarm P_i using Eq.(2);
  end
end

III. EXPERIMENTAL STUDIES

A. Experimental setup

Seven benchmark functions proposed in CEC'08 special session on large scale optimization [18] are used in this section to test QPSO_CC and compare with other algorithms (as listed in Table 1), in which f_1, f_4 and f_6 are separable functions, f_2 , f_3 , f_5 and f_7 are non-separable functions.

Experiments are conducted on the above 7 functions for 100, 500, and 1000 dimensions. For each test function, the average results of 25 independent runs were recorded. For each run, maximum number of fitness evaluations (Max_FES) were set to 5000* D. A two-tailed t-test was conducted with a null hypothesis stating that there is no difference between two algorithms in comparison. The population size for each swarm involved in coevolution is set to 30.

Experiments are implemented on a Server with an Intel Xeon CPU E7-4809 (4 processors) and 128 GB RAM. The algorithm is written in MATLAB on matlab 2011b by using the parallel computing toolbox.

B. Results and analysis

First, to verify the performance of QPSO_CC for large scale optimization, comparison with two of the state-of-the-art algorithms (CSO [2], CCPSO2 [14]) is given in Table 2, 3, and 4. The same criteria proposed in [18] is adopted here.

From the results, it is noted that QPSO_CC shows better performance as the dimension size increases.

TABLE 2. RESULTS OF QPSO_CC, CCPSO2 AND CSO ON TEST FUNCTIONS OF 100 DIMENSIONS

100-D	Mean best fitness (standard deviation)		
	CSO	CCPSO2	QPSO_CC
f_1	9.02E-15 (5.53E-15)	7.73E-14 (3.23E-14)	9.11E-29 (1.10E-28)
f_2	2.31E+01 (1.39E+01)	6.08E+00 (7.83E+00)	3.35E+01 (5.38E+00)
f_3	4.31E+00 (1.26E+01)	4.23E+02 (8.65E+02)	3.90E+02 (5.53E+02)
f_4	2.78E+02 (3.43E+01)	3.98E-02 (1.99E-01)	5.60E+01 (7.48E+00)
f_5	2.96E-04 (1.48E-03)	3.45E-03 (4.88E-03)	0 (0)
f_6	2.12E+01 (4.02E-01)	1.44E-13 (3.06E-14)	1.20E-14 (1.52E-15)
f_7	-1.39E+03 (2.64E+01)	-1.50E+03 (1.04E+01)	-7.28E+05 (1.88E+04)

Table 3. Results of QPSO_CC, CCPSO2 and CSO on test functions of 500 dimensions

500-D	Mean best fitness (standard deviation)		
	CSO	CCPSO2	QPSO_CC
f_1	2.25E-14 (6.10E-15)	3.00E-13 (7.96E-14)	6.75E-23 (3.90E-24)
f_2	2.12E+01 (1.74E+01)	5.79E+01 (4.21E+01)	2.60E+01 (2.4E+00)
f_3	2.93E+02 (3.59E+01)	7.24E+02 (1.54E+02)	5.74E+02 (1.67E+02)
f_4	2.18E+03 (1.51E+02)	3.98E-02 (1.99E-01)	3.19E+02 (2.16E+01)
f_5	7.88E-04 (2.82E-03)	1.18E-03 (4.61E-03)	2.22E-16 (0.00E+00)
f_6	2.15E+01 (3.10E-03)	5.34E-13 (8.61E-14)	4.13E-13 (1.10E-14)
f_7	-6.37E+03 (7.59E+01)	-7.23E+03 (4.61E+01)	-1.97E+06 (4.08E+04)

Table 4. Results of QPSO_CC, CCPSO2 and CSO on test functions of 1000 dimensions

1000-D	Mean best fitness (standard deviation)		
	CSO	CCPSO2	QPSO_CC
f_1	7.81E-15 (1.52E-15)	5.18E-13 (9.61E-14)	1.09E-21 (4.20E-23)
f_2	3.65E+02 (9.02E+00)	7.82E+01 (4.25E+01)	4.15E+01 (9.74E-01)
f_3	9.10E+02 (4.54E+01)	1.33E+03 (2.63E+02)	1.01E+03 (3.02E+01)
f_4	5.31E+03 (2.48E+02)	1.99E-01 (4.06E-01)	6.89E+02 (3.10E+01)
f_5	3.94E-04 (1.97E-03)	1.18E-03 (3.27E-03)	2.26E-16 (2.18E-17)
f_6	2.15E+01 (3.19E-01)	1.02E-12 (1.68E-13)	1.21E-12 (2.64E-14)
f_7	-1.25E+04 (9.36E+01)	-1.43E+04 (8.27E+01)	-3.83E+06 (4.82E+04)

IV. CONCLUSIONS

In this paper, QPSO with cooperative coevolution (QPSO_CC) is used to decompose the high dimensional problems into several lower dimensional problems and optimize them separately. The numerical experimental results show that QPSO_CC has comparative or even better performance than other algorithms.

ACKNOWLEDGMENT

This work is supported by Jiangsu Postdoctoral Funding (Project No. 1401004B).

REFERENCES

[1] Ran Cheng, Yaochu Jin, A social learning particle swarm optimization, Information Sciences, 291, 43-60, 2015.

[2] Ran Cheng, Yaochu Jin, A competitive swarm optimizer for large scale optimization, IEEE Transaction on Cybernetics, 45(2), 191-204, 2015.

[3] Frans van den Bergh, Andries P. Engelbrecht, A cooperative approach to particle swarm optimization, IEEE Transaction on Evolutionary Computation, 8(3) 225-239, 2004.

[4] MARCO A. Montes de Oca, Ken Van den Enden, Marco Dorigo, Incremental social learning in particle swarms, IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, 41(2), 368-384, 2011.

[5] Zhenyu Yang, Ke Tang and Xin Yao, Large scale evolutionary optimization using cooperative coevolution, Information Sciences, 178, 2985-2999, 2008.

[6] Zhi-Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams, Big data opportunities and challenges: discussions from data analytics perspectives, IEEE Computational Intelligence Magazine, 2014.

[7] O. Olorunda, A. Engelbrecht, Measuring exploration/exploitation in particle swarms using swarm diversity, Proceeding of IEEE Congress on Evolutionary Computation, 2008, 1128-1134.

[8] A. Ismail, A. Engelbrecht, Measuring diversity in the cooperative particle swarm optimizer, Swarm Intelligence, 7461, 97-108, 2012.

[9] N. Tian, C.H. Lai, Parallel quantum-behaved particle swarm optimization, International Journal of Machine Learning and Cybernetics, 5(2), 309-318, 2014.

[10] M.A. Potter, K.A. de Jong, A cooperative coevolutionary approach to function optimization, The Third Parallel Problems Solving From Nature, Berlin, Germany: Springer-Verlag, 1994, 249-257.

[11] Frans van den Bergh, A.P. Engelbrecht, A cooperative approach to particle swarm optimization, IEEE Transactions on evolutionary computation, 8(3), 225-239, 2004.

[12] Z. Yang, K. Tang and X. Yao, Large scale evolutionary optimization using cooperative coevolution, Information Science, 178(15), 2986-2999, 2008.

[13] M. Omidvar, Cooperative co-evolution for large scale optimization through more frequent random grouping, Proceeding of CEC, 2010, 1754-1761.

[14] X.D. Li, X. Yao, Cooperatively coevolving particle swarms for large scale optimization, IEEE Transactions on evolutionary computation, 16(2), 210-214, 2012.

[15] J. Sun, B. Feng and W.B. Xu, Particle swarm optimization with particles having quantum behaviour, IEEE Congress Evolutionary Computation, Portland, USA, 2004.

[16] J. Sun, Quantum-behaved particle swarm optimization: analysis of individual particle behavior and parameter selection, Evolutionary Computation, 20(3), 349-393, 2012.

[17] J. Sun, Convergence analysis and improvements of quantum-behaved particle swarm optimization, Information Sciences, 193, 81-103, 2012.

[18] K. Tang, X. Yao, P. Suganthan, Benchmark functions for the CEC'08 special session and competition on large scale global optimization, Nature Inspired Computation and Applications Laboratory, USTC, China, 2007. (http://nical.ustc.edu.cn/cec08ss.php).

TABLE 1. BENCHMARK FUNCTIONS FOR LARGE SCALE GLOBAL OPTIMIZATION

function	definition	domain
f_1	$f_1(X)=\sum_{i=1}^{D} z_i^2$	$[-100,100]^D$
f_2	$f_2(X)=\max\{\lvert z_i\rvert, 1\le i\le D\}$	$[-100,100]^D$
f_3	$f_3(X)=\sum_{i=1}^{D-1}\left(100\left(z_i^2-z_{i+1}\right)^2+\left(z_i-1\right)^2\right)$	$[-100,100]^D$
f_4	$f_4(X)=\sum_{i=1}^{D}\left(z_i^2-10\cos(2\pi z_i)+10\right)$	$[-5,5]^D$
f_5	$f_5(X)=\sum_{i=1}^{D}\frac{z_i^2}{4000}-\prod_{i=1}^{D}\cos\left(\frac{z_i}{\sqrt{i}}\right)+1$	$[-600,600]^D$
f_6	$f_6(X)=-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} z_i^2}\right)$ $-\exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos\left(2\pi z_i\right)\right)+20+e$	$[-32,32]^D$
f_7	$f_7(X)=\sum_{i=1}^{D} fractal1D\left(x_i+twist(x_{(i \bmod D+1)})\right)$ $twist(y)=4\left(y^4-2y^3+y^2\right)$ $fractal1d(x)\approx$ $\sum_{k=1}^{3}\sum_{1}^{2^k-1}\sum_{1}^{ran2(o)} dip\left(x, ran1(0), \frac{1}{2^{k-1}(2-ran1(o))}\right)$ $dip(x,c,s)=\begin{cases}\left(-6144(x-c)^6+3088(x-c)^4-392(x-c)^2+1\right)s, & -0.5<x<0.5\\ 0, \text{otherwise}\end{cases}$	$[-1,1]^D$

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Path planning for welding spot detection

Xia Zhu

State Key Laboratory of Mechanics and Control of Mechanical Structures,
Nanjing University of Aeronautics & Astronautics,
Nanjing, China
Email: snail2024@163.com

Renwen Chen

State Key Laboratory of Mechanics and Control of Mechanical Structures,
Nanjing University of Aeronautics & Astronautics,
Nanjing, China
Email: rwchen@nuaa.edu.cn

***Abstract*— Aiming at the path optimization of welding spot detection, a novel particle swarm optimization algorithm was proposed and adjusted the parameter in combination with the neighborhood information, which can overcome the local minimum problem to some degree. In the process of path planning, the minimum time was taken as movement criterion. At the current path point, a path planning time function was established. By using the novel particle swarm algorithm, the minimum point of the time function was obtained and taken as the next path point, thus the path of the welding spot could be acquired step by step. The simulation results showed that an effective optimal path was planned by applying the proposed method, which could fulfill the task of welding spot traversal. Finally the theoretical analysis results were verified with simulations, and it was shown that the new algorithm has better convergence accuracy and controllable speed compared with other PSO algorithms.**

keywords— welding spot; path planning; test function; particle swarm optimization

1. INTRODUCTIONS

The surface area of components on printed circuit boards (PCB) is becoming smaller and smaller, and the pins of components is more and more intensive, which make welding spot detection more and more difficult. Welding spot can be detected with image processing method, but how to move through numerous welding spots within the shortest time is also a worth studying problem.If we take every spot on PCB as a city, then the path of the welding spot detection is equal to path planning, similar to the TSP problem.The TSP problem always has an important theoretical research value and a wide range of engineering application. Path planning algorithm, such as branch and bound method, dynamic programming and so on, can get the optimal solution in smaller PCB size. But with the increase of the scale, time and space complexities of the algorithm will be significantly increased.

Optimization algorithm has strong search ability, while the path planning is to design an algorithm to obtain an optimal path, therefore path planning problem is an important branch of the optimization algorithm research. With the development of computer technology, some new optimization algorithms have been proposed, including the ant algorithm, genetic algorithm, particle swarm optimization and so on. Ant algorithm is first proposed to solve TSP problem, but the basic ant algorithm has slow convergence speed and easily fall into local optimum, and the values of the parameters in the model are directly related to the convergence speed and global search ability. Genetic algorithm has the following main disadvantages: it is more complex in programming; the select of parameters relies on experience; its search speed is slow; it falls easily into "premature" for complex combinatorial optimization problem when the search space is large. Particle swarm optimization algorithm has been highly concerned for the following advantages: strong commonality, simple principle, easy to implement and independent on the problem. But the basic particle swarm optimization traps into local minimum solution easily, and has poor local search ability, low computational efficiency and poor antinoise ability.

Therefore, in view of the slow convergence speed and the defects of large amount of calculation in the later stage of basic particle swarm algorithm, we put forward a novel algorithm and applies it to solve welding spot path planning problem in this paper. Local impact factor, which gradually increases with the increase of iterations, is taken into the algorithm to reduce calculation amount and speeds up the convergence. Experimental results show that the welding spot detection precision and speed have improved after using this algorithm, which has significant practical value.

2. PROBLEM DESCRIPTION

Welding spots on PCB have different regions, different directions and different lengths between each other. Path planning problem refers to the calculation of the shortest distance from the starting point to the destination. Assuming that the current welding spot is P, we calculate the minimum distance according to the following steps: Start from point P, probe in all directions, calculate the distances between the starting point and the candidate points and select the shortest distance, as shown in Fig.1. The search of the next welding spot is constraint by angular velocity and angle. θ is the angle, and $\theta_{\max}$ is the maximum angle; ϖ is the angular velocity and $\varpi_{\max}$ is the maximum angular velocity; then $|\theta| \le \theta_{\max}$, $|\varpi| \le \varpi_{\max}$. $p_0(x_0, y_0)$ is the starting welding spot, the welding spots on the way are $p_i(x_i, y_i)$, where $i = 1, 2, ..., num\text{-}1$, and $p_{num}(x_{num}, y_{num})$ is the last welding spot.

In each search, the main target is to find the optimal θ and ϖ. A path planning time function should be established before each search, then an improved particle swarm optimization algorithm is applied to find an optimal point as the next starting point in the search field.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.29

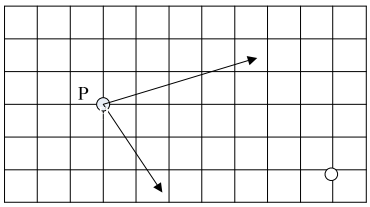


Fig.1 Search length and angle calculation area

To minimize the total detection time of the detected PCB shown in Fig.2, an optimal path and a time sequence need to be selected.

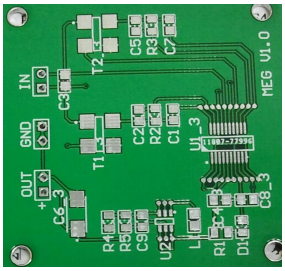


Fig.2 The detected PCB

(1) Movement time

Assuming the distance between current spot p_i and candidate spot p_j is d_{ij}, the time for moving in per unit length is α, and then the total time for completing all num spots is:

$$T1=\sum_{i=1}^{num}\sum_{j=1,j\neq i}^{num}\alpha\times d_{ij}\times\lambda_{ij} \quad (1)$$

$$\lambda_{ij}=\frac{1}{\sqrt{2\pi}}\exp[-\frac{(\theta_{ij}-\theta_0)^2}{2\sigma_1^2}-\frac{(h_{ij}-h_0)^2}{2\sigma_2^2}] \quad (2)$$

Where, the angle and the horizontal distance between current spot p_i and destination spot p_{num} is (θ_0,h_0), the angle and the horizontal distance between current spot p_i and candidate spot p_j is (θ_{ij},h_{ij}). When spot p_i moves to p_j, the closer (θ_{ij},h_{ij}) is to (θ_0,h_0), the larger the relation function λ_{ij} is, and vice versa. d_{ij} can be obtained according to the following equation besides through Euler's formula:

$$d_{ij}=h_{ij}/\cos(\theta_{ij}) \quad (3)$$

Thus movement time function can rewrite as the following equation:

$$T1=\sum_{i=1}^{num}\sum_{j=1,j\neq i}^{num}\alpha\times(h_{ij}/\cos(\theta_{ij}))$$

$$\times\frac{1}{\sqrt{2\pi}}\exp[-\frac{(\theta_{ij}-\theta_0)^2}{2\sigma_1^2}-\frac{(h_{ij}-h_0)^2}{2\sigma_2^2}] \quad (4)$$

(2) Testing time

Welding spots can be segmented from the background through image processing. Each spot has a corresponding coordinates, area, shape and so on. Testing time $T2$ is set according to the following equation:

$$T2=\sum_{i=1}^{num}(\kappa_i+\mu_i) \quad (5)$$

Where, κ_i is the time for searching all welding spots from PCB, μ_i is the time for detecting the weld.

To solve the minimum of $T1+T2$, namely,

$$\min(T1+T2) \quad (6)$$

In conclusion, welding spot path planning problem is converted into finding a combination of num spots to minimize $T1+T2$. This paper focuses on path planning, so $T2$ is assumed to be constant here. During path planning, the minimum point will be found in the search field by using the improved particle swarm algorithm, which will lead to finding the next point.

3. MODEL OF THE NOVEL PARTICLE SWARM OPTIMIZATION ALGORITHM

3.1 particle swarm optimization

Inspired by social behavior of bird flocking or fish schooling, Dr. Eberhart and Dr. Kennedy proposed particle swarm optimization in 1995. Particle swarm optimization algorithm is simple, robust and easy to implement. This algorithm belongs ideologically to that philosophically school that allows wisdom to emerge rather than trying to impose it, that emulates nature rather than trying to control it, and that seeks to make things simpler rather than more complex[8]. Since this algorithm is proposed, the domestic and foreign scholars have paid many attentions in order to improve the performance through thorough theoretical analysis, and put forward some improvements. At present, improvements are concentrated in the following aspects: the position and velocity updating formula, setting many species, adding other intelligent into particles, the topology of the group, hybrid methods and so on. The typical algorithms include PSO, AIWFSO, SPSO ,etc.

1. The basic particle swarm optimization (PSO)

$$v_i(t+1)=v_i(t)+c_1*rand_1()*(p_ibest-x_i(t))$$
$$+c_2*rand_2()*(g_ibest-x_i(t)) \quad (7)$$

$$x_i(t+1)=x_i(t)+v_i(t+1) \quad (8)$$

Where, $rand_k()\sim U(0,1)$, $k=1,2$, c_1 and c_2 are acceleration coefficients and normally are set to 2, $v_i(t)$ is the velocity of particle i at time t, $x_i(t)$ is the position

of particle i at time t, $p_i best$ is the personal best solution of particle i at time t, $g_i best$ is the best position found by the neighborhood of particle i at time t.

2. Particle swarm optimization with adaptive inertia weight factor (AIWFPSO)

ω is the weight coefficient and the key to affect search behavior. In the process of solution space optimization, PSO algorithm is a kind of nonlinear operation. In order to balance global optimization and local optimization abilities, particle swarm optimization algorithm with adaptive inertia weight factor was put forward, in which ω changes with the fitness function value automatically.

$$v_i(t+1) = \omega v_i(t) + c_1 * rand_1() * (p_i best - x_i(t)) + c_2 * rand_2() * (g_i best - x_i(t)) \quad (9)$$

$$\omega = \begin{cases} \omega_{\min} + \dfrac{(\omega_{\max} - \omega_{\min})(f - f_{\min})}{f_{avg} - f_{\min}}, f \le f_{avg} \\ \omega_{\max}, f \ge f_{avg} \end{cases} \quad (10)$$

Where, $\omega_{\min}$ and $\omega_{\max}$ represent the minimum and maximum of ω respectively, $f_{\min}$ and f_{avg} are the minimum fitness function value and the average fitness function value of all particles respectively, f is the current fitness function value.

3. Particle swarm optimization based on group strategy information (SPSO)

In the basic particle swarm optimization algorithm, the following information is already known: the historical optimal position of a group and its corresponding fitness function value, the historical position of any individual particle and its corresponding fitness function value. Group strategy gets the next position to guide the update of the particle according to the above information. Assume that $p_j(t)$ is the individual historical position of particle j, the fitness function value of corresponding position can be used as its weight. For particle j, the weight of its individual historical position is set equal to π_j, and $\sum_{j=1}^{m} \pi_j = 1$, then we can get a better position $p_{gd}(t)$ by using group strategy. Thus the evolution equation of the particle swarm algorithm based on group strategy can be obtained as below:

$$v_i(t+1) = \omega * v_i(t) + c_1 * rand_1() * (p_i best(t) - x_i(t)) + c_2 * rand_2() * (p_{gd}(t) - x_i(t)) \quad (11)$$

Where,

$$p_{gd}(t) = \sum_{j=1}^{m} \pi_j p_j best(t) \quad (12)$$

$$\pi_j = \frac{e^{score_j(t)}}{\sum_{u=1}^{m} e^{score_u(t)}} \quad (13)$$

$$score_j(t) = \begin{cases} 1, f_{worst}(p_j(s)) = f_{best}(p_j(s)) \\ \dfrac{f_{worst}(p_j(s)) - p_j(t)}{f_{worst}(p_j(s)) - f_{best}(p_j(s))} \end{cases} \quad (14)$$

$$f_{worst}(p_j(s)) = \arg\max\{f(p_k(t)) \mid k = 1,2,...,m\} \quad (15)$$

$$f_{best}(p_j(s)) = \arg\min\{f(p_k(t)) \mid k = 1,2,...,m\} \quad (16)$$

4. A novel particle swarm optimization (ZPSO)

According to the above theoretical analysis, a novel particle swarm algorithm was proposed in this paper. Neighborhood information was added to particle velocity updating equation. In each iteration, neighborhood optimal value and global optimal value can be adjusted dynamically so that the optimization is carried out mainly by global optimization, supplemented by neighborhood optimization.

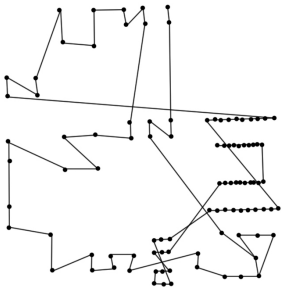
The particle velocity updating equation in the novel particle swarm optimization is as follows:

$$v_i(t+1) = \omega * v_i(t) + c_1 * rand_1() * (p_i best(t) - x_i(t)) + \frac{c_2 * rand_2() * (g_i best(t) - x_i(t))}{t} + \frac{c_3 * rand_3() * (l_i best(t) - x_i(t)) * (t-1)}{t} \quad (17)$$

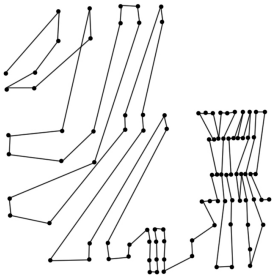
Where, $p_i lbest(t)$ is the optimal value of neighborhood, $rand_3() \sim U(0,1)$, c_3 is acceleration coefficient and the rest parameters are defined as PSO. In PSO algorithm, whether it can find the optimal solution efficiently depends on how to balance the global and local search abilities. Balance effect can be determined by adjusting the parameters of PSO. Compared to Equation (7), the neighborhood information was added into Equation (16), and the time factor was added in the last two items, which showed that global optimal information and the optimal neighborhood information could dynamically update over time. From Equation (16) ,we can see that when t is small, individual particles depend more on the global optimal information, but with the increase of time, they become more dependent on the neighborhood optimal information to avoid the premature of particles so that this algorithm has better global and local optimization ability.

4. PATH PLANNING SIMULATION EXPERIMENT

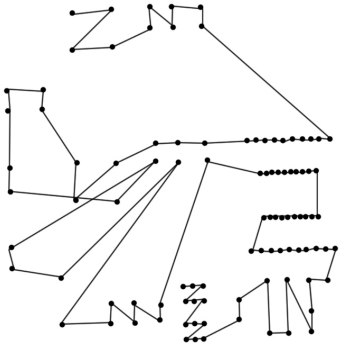
Take the path planning problem of PCB shown in Fig.2 for instance. We compared ZPSO with PSO, AIWFPSO, SPSO, CPSO, etc in the minimization of time cost. Figure 4 showed the algorithm performances in path planning optimization. According to Figure 4, ZPSO algorithm had a better ability to avoid repeated steps and the waste of time, which means that ZPSO could effectively solve the path planning problem.



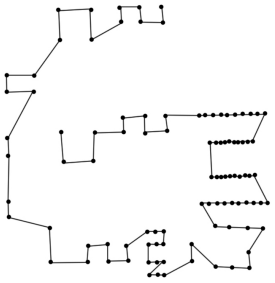
(a) path planning based on PSO



(b) path planning based on AIWFPSO



(c) path planning based on SPSO



(d) path planning based on ZPSO

Fig. 4 The path planning comparison under different algorithms

5. CONCLUSIONS

In this paper, the movement of welding spot was transformed into path planning problem. Firstly, a mathematical model of path planning problem was established. Then ZPSO algorithm was introduced in order to optimize the fitness function, find the minimum point and get optimal solution.

This paper has expounded and proved that the ZPSO algorithm has better convergence compared to other PSO algorithms. PSO has its advantages as a new optimization algorithm which attracts more researchers' interests.But its foundation is not perfect and easily leads to local minimum which need to be improved. In this paper, we propose a novel PSO with the neighborhood information, which can overcome the local minimum problem to some degree, and the application performance is very good.

The simulation results showed that ZPSO algorithm can better avoid repeated step and the waste of time and eventually get a highly efficient path of welding spot.

ACKNOWLEDGMENT

The authors thank the reviewers and editors for their corresponding contributions in making the paper more presentable. This research is supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

REFERENCE

1. Bentner J. et al., Optimization of the time-dependent traveling salesman problem with Monte Carlo methods, Physical Review E, Vol. 64, 036701, 2001.
2. Hameed A. An optimized field coverage planning approach for navigation of agricultural robots in fields involving obstacle areas. International Journal of Advanced Robotic Systems , Vol: 10 (231) ,pp. 1-9,2013
3. Curiac, DI . A 2D chaotic path planning for mobile robots accomplishing boundary surveillance missions in adversarial conditions. COMMUNICATIONS IN NONLINEAR SCIENCE AND NUMERICAL SIMULATION,vol 19(10) 3617-3627,2014
4. J. Kennedy and R. C. Eberhart "Particle Swarm Optimization", Proc. IEEE International Conference on Neural Networks, vol. IV, pp.1942 -1948 1995.
5. R. C. Eberhart and J. Kennedy "A New Optimizer Using Particles Swarm Theory", Proc. Sixth International Symposium on Micro Machine and Human Science (Nagoya, Japan), pp.39 -43 1995 .
6. Shi Y, Eberhart R C. A modified particle swarm optimizer. In: Proc. of the IEEE CEC.1998:69-73.
7. XIAO Renbin. Swarm intelligence for complex systems [M]. Beijing: science press, 2013.
8. Bentner J. et al., Optimization of the time-dependent traveling salesman problem with Monte Carlo methods, Physical Review E, Vol. 64, 036701, 2001.
9. Clerc M., Kennedy J., The particle swarm-explosion, stability, and convergence in a multidimensional complex space, IEEE Transactions on Evolutionary Computation, Vol. 6, pp. 58-73, 2002.
10. Kennedy J., Eberhart R.C., Discrete binary version of the particle swarm algorithm, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, pp. 4104-4108, 1997.
11. Yan Zheping, DENG chao, ZHAO Yufei etc. path planning method for UUV near sea bottom [J]. Journal of Harbin engineering university. 2014, 35 (3) : 307-312
12. Spears, W. M., Green, D. T. & Spears, D. F.. Biases in Particle Swarm Optimization. International Journal of Swarm Intelligence Research, Vol. 1(2), pp. 34-57,2010

Characters of a Class of a Rational Difference Equation $x_{n+1}=\frac{ax_nx_{n-1}}{bx_{n-1}-cx_n}$

XIAO Qian
Department of Basic Courses
Hebei Finance University
Baoding 071000, China
xiaoxiao_xq168@163.com

WANG Li-bin
Department of Basic Courses
Hebei Finance University
Baoding 071000, China
libinwangtju@126.com

TANG Jie
Human Resources Department
State Grid Baoding Electric Power Supply Company
Baoding 071000, China
2013756858@qq.com

***Abstract*: This paper is concerned with the following rational difference equation $x_{n+1}=ax_nx_{n-1}/\left(bx_{n-1}-cx_n\right)$, with the initial conditions $x_{-1},x_0\in(0,+\infty)$, and $a,b,c\in R^+$. Locally asymptotically stability, global attractivity and boundedness character of the equilibrium point of the equation are investigated. Moreover, simulation is shown to support the results.**

Keyword:Global stability, Attractivity, Boundedness, Numerical simulation

I. INTRODUCTION

Difference equations are applied in the field of biology, engineer, physics and so on[1]. The study of properties of rational difference equations have been an area of intense interest in recent years. There has been a lot of work[5-8] dealing with the qualitative behavior of rational difference equation. For example, Agarwal et al. [2] investigated the global stability, periodicity character and gave the solution of some special cases of the difference equation

$$x_{n+1}=a+\frac{dx_{n-l}x_{n-k}}{b-cx_{n-5}}.$$

Saleh et al. [3,4] studied the difference equation

$$y_{n+1}=A+\frac{y_n}{y_{n-k}}.$$

In [5] C. Wang et al. dealt with the asymptotic behavior of equilibrium point for the rational difference equation

$$x_{n+1}=\frac{\sum_{i=1}^{l}A_{s_i}x_{n-s_i}}{B+C\prod_{j=1}^{k}x_{n-t_j}}.$$

In [6] Elabbasy et al. has got the global stability, periodicity character and gave the solution of special case of the following recursive sequence

$$x_{n+1}=ax_n-\frac{bx_n}{cx_n-dx_{n-1}}.$$

In this paper we consider the qualitative behavior of rational difference equation

$$x_{n+1}=\frac{ax_nx_{n-1}}{bx_{n-1}-cx_n},\ n=0,1,\cdots \tag{1}$$

with initial data $x_{-1},x_0\in(0,+\infty)$, and $a,b,c\in R^+$.

II. PRELIMINARIES AND NOTATION

Let us introduce some basic definitions and some theorems that we need in what follows.

Lemma 1 Let I be some interval of real numbers and

$$f:I^{k+1}\to I$$

be a continuously differentiable function. Then for every set of initial conditions $x_{-k},x_{-k+1},\cdots,x_0\in I$, the difference equation

$$x_{n+1}=f(x_n,x_{n-1},\cdots,x_{n-k}),\ n=0,1,\cdots \tag{2}$$

has a unique solution $\{x_n\}_{n=-k}^{\infty}$.

Definition 1(Equilibrium point) A point $\bar{x}\in I$ is called an equilibrium point of (2), if

$$\bar{x}=f(\bar{x},\bar{x},\cdots,\bar{x})$$

Definition2 (Stability)（1）The equilibrium point $\bar{x}$ of (2) is locally stable if for every $\varepsilon>0$, there exists $\delta>0$, such that for any initial data $x_{-k},x_{-k+1},\cdots,x_0\in I$, with

$$\left|x_{-k}-\bar{x}\right|+\left|x_{-k+1}-\bar{x}\right|+\cdots+\left|x_0-\bar{x}\right|<\delta$$

we have $\left|x_n-\bar{x}\right|<\varepsilon$, for all $n\ge -k$.

(2) The equilibrium point $\bar{x}$ of (2) is locally asymptotically stable if $\bar{x}$ is locally stable solution of (2), and there exists $\gamma>0$, such that for all $x_{-k},x_{-k+1},\cdots,x_0\in I$, with

$$\left|x_{-k}-\bar{x}\right|+\left|x_{-k+1}-\bar{x}\right|+\cdots+\left|x_0-\bar{x}\right|<\gamma$$

we have

$$\lim_{n\to\infty}x_n=\bar{x}.$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.30

(3) The equilibrium point $\bar{x}$ of (2) is global attractor if for all $x_{-k}, x_{-k+1}, \cdots, x_0 \in I$, we have $\lim_{n\to\infty} x_n = \bar{x}$.

(4) The equilibrium point $\bar{x}$ of (2) is globally asymptotically stable if $\bar{x}$ is locally stable and $\bar{x}$ is also global attractor of (2).

(5) The equilibrium point $\bar{x}$ of (2) is unstable if $\bar{x}$ is not locally stable.

Definition 3 The linearized equation of (2) about the equilibrium $\bar{x}$ is the linear difference equation

$$y_{n+1} = \sum_{i=0}^{k} \frac{\partial f(\bar{x}, \bar{x}, \cdots, \bar{x})}{\partial x_{n-i}} y_{n-i} \tag{3}$$

Lemma 2[2] Assume that $p_1, p_2, \cdots, p_k \in R$ and $k \in \{1, 2, \cdots\}$, then

$$\sum_{i=1}^{k} |p_i| < 1$$

is a sufficient condition for the asymptotic stability of the difference equation

$$x_{n+k} + p_1 x_{n+k-1} + \cdots + p_k x_n = 0,\ n = 0, 1, \cdots \tag{4}$$

Lemma 3[1] Let $g : [p,q]^2 \to [p,q]$ be a continuous function, where p and q are real numbers with $p < q$ and consider the following equation

$$x_{n+1} = g(x_n, x_{n-1}),\ n = 0, 1, \cdots \tag{5}$$

Suppose that g satisfies the following conditions:

(1) $g(x, y)$ is non-decreasing in $x \in [p,q]$ for each fixed $y \in [p,q]$, and $g(x, y)$ is non-increasing in $y \in [p,q]$ for each fixed $x \in [p,q]$.

(2) If (m, M) is a solution of system

$$M = g(M, m) \text{ and } m = g(m, M)$$

then $M = m$.

Then there exists exactly one equilibrium $\bar{x}$ of (5), and every solution of (5) converges to $\bar{x}$.

III. THE MAIN RESULTS AND THEIR PROOFS

In this section we investigate the local stability character of the equilibrium point of (1).

If $b - c \neq a$, then the unique equilibrium point of equation (1) is $\bar{x} = 0$.

Let $f : [0, \infty)^2 \to [0, \infty)$ be a function defined by

$$f(u, v) = \frac{auv}{bv - cu} \tag{6}$$

Therefore it follows that

$$f_u(u, v) = \frac{abv^2}{(bv - cu)^2},\ f_v(u, v) = -\frac{acu^2}{(bv - cu)^2}.$$

Theorem 1 Assume that $ab + ac < (b - c)^2$, then the equilibrium point $\bar{x} = 0$ of (1) is locally asymptotically stable.

Proof. When $\bar{x} = 0$,

$$f_u(\bar{x}, \bar{x}) = \frac{ab}{(b - c)^2},\quad f_v(\bar{x}, \bar{x}) = -\frac{ac}{(b - c)^2}$$

The linearized equation of (1) about $\bar{x} = 0$ is

$$y_{n+1} - \frac{ab}{(b - c)^2} y_n + \frac{ac}{(b - c)^2} y_{n-1} = 0 \tag{7}$$

It follows by Lemma 2, (7) is asymptotically stable, if

$$\left| \frac{ab}{(b - c)^2} \right| + \left| \frac{ac}{(b - c)^2} \right| < 1$$

or

$$\frac{ab + ac}{(b - c)^2} < 1$$

so

$$ab + ac < (b - c)^2$$

This completes the proof.

Theorem 2 The equilibrium point $\bar{x} = 0$ of (1) is global attractor if $c \neq 0$.

Proof. Let p, q are real numbers and assume that $g : [p,q]^2 \to [p,q]$ is a function defined by $g(u, v) = \frac{auv}{bv - cu}$, then we can easily see that the function $g(u, v)$ is increasing in u and is decreasing in v.

Suppose that (m, M) is a solution of system

$$M = g(M, m) \text{ and } m = g(m, M).$$

Then from (1)

$$M = \frac{aMm}{bm - cM},\ m = \frac{amM}{bM - cm}.$$

Therefore

$$bMm - cM^2 = aMm, \tag{8}$$

$$bMm - cm^2 = aMm. \tag{9}$$

Subtracting (9) from (8) gives

$$c\left(M^2 - m^2\right) = 0.$$

Since $c \neq 0$, it follows that

$$M = m.$$

Lemma 3 suggests that $\bar{x}$ is global attractor of (1) and then the proof is completed.

Theorem 3 For initial data x_{-1}, x_0, if $a \leq c$, that every solution x_n of (1) is bounded, and $x_n \leq M$, here $M = \max\left\{x_{-1}, x_0\right\}$.

Proof. Let $\left\{x_n\right\}_{n=-1}^{\infty}$ be a solution sequence of (1). It follows from (1) that

$$x_{n+1} = \frac{ax_n^2}{cx_n + bx_{n-1}} \leq \frac{ax_n^2}{cx_n} = \frac{ax_n}{c}$$

Then $x_{n+1} \leq x_n$ for all $n \geq 0$.

Then the sequence $\left\{x_n\right\}_{n=0}^{\infty}$ is decreasing and bounded derived from $M = \max\left\{x_{-1}, x_o\right\}$.

Corollary 1 For any initial data x_{-1}, x_0, if $c \neq 0$ and $ab + ac < \left(b - c\right)^2$, then the equilibrium point $\bar{x} = 0$ of (1) is globally asymptotically stable.

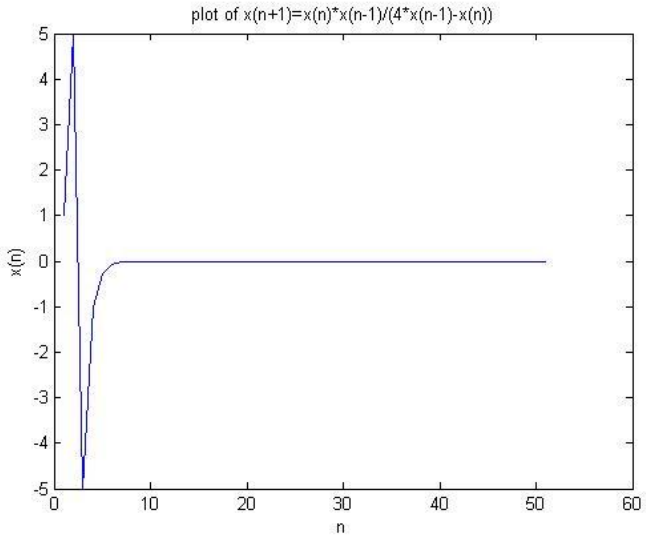
IV. NUMERICAL SIMULATION

In this section, we give some numerical simulations to support our theoretical analysis. For example, we consider the equation

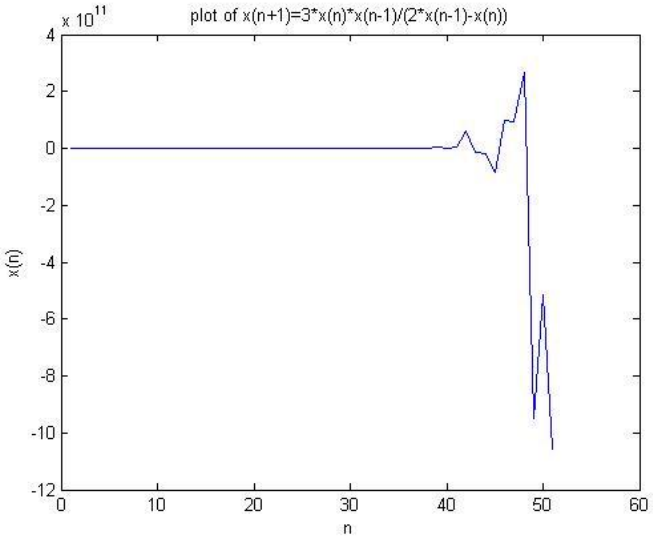
$$x_{n+1} = \frac{x_n x_{n-1}}{4x_{n-1} - x_n} \tag{10}$$

$$x_{n+1} = \frac{3x_n x_{n-1}}{2x_{n\ 1} - x_n} \tag{11}$$

We can present the numerical solutions of (10) and (11) which are shown, respectively in Figure 1 and 2. Figure 1 shows the asymptotic behavior and boundedness of the solution to (10) with initial data $x_0 = 1,\ x_1 = 5$, Figure 2 shows the dispersed behavior of the solution to (11) with initial data $x_0 = 1,\ x_1 = 3$.



Figue 1 This figure shows the solution of (10), where $x_0 = 1,\ x_1 = 5$



Figue 2 This figure shows the solution of (10), where $x_0 = 1,\ x_1 = 3$

V. ACKNOWLEDGMENTS

The authors would like to thank the reviewers and the editor for their valuable suggestions and comments. This work is supported in part by Applied Mathematics outstanding basic subjects in Hebei Finance University.

REFERENCES

[1] L. Berezansky, E. Braverman, E. Liz, Suffcient conditions for the global stability of nonautonomous higher order difference equations, J. Difference Equ.Appl. 11 (9) 2005, pp. 785–798.

[2] R.P. Agarwal, E.M. Elsayed, Periodicity and stability of solutions of higher order rational difference equation, Adv. Stud. Contemp. Math. 17 (2) 2008, pp. 181-201.

[3] M. Saleh, M. Aloqeili, On the difference equation $y_{n+1} = A + y_n / y_{n-k}$ with $A < 0$, Appl. Math. Comput. 176 (1) 2006, pp. 359-363.

[4] M. Saleh, M. Aloqeili, On the difference equation $x_{n+1} = A + x_n / x_{n-k}$, Appl. Math. Comput. 171, 2005, pp. 862-869.

[5] C. Wang, Q. shi, S.Wang, Asymptotic behavior of equilibrium point for a family of rational difference equation. Advances in Difference Equations. 2010. Article ID 505906.

[6] E.M. Elabbasy, H. El-Metwally, E.M. Elsayed, On the difference equation $x_{n+1} = ax_n - bx_n / \left(cx_n - dx_{n-1}\right)$, Adv. Difference Equ. 2006 , pp. 1-10. Article ID 82579.

[7] E.M.E. Zayed. M.A. El-Moneam, On the rational recursive sequence $x_{n+1} = \left(\alpha x_n + \beta x_{n-1} + \gamma x_{n-2} + \delta x_{n-3}\right) / \left(Ax_n + Bx_{n-1} + Cx_{n-2} + Dx_{n-3}\right)$, Comm. Appl. Nonlinear Anal. 12, 2005, pp. 15-28.

[8] R. Memarbashi, Sufficient conditions for the exponential stability of nonautonomous difference equations, Applied Mathematics Letters, 3 (21), 2008, pp. 232–235.

A Novel Quantum-behaved Particle Swarm Optimization Algorithm

Jing Zhao[1,2]
[1]School of Information, Qilu University of Technology
[2]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Shandong Normal University
Jinan, China, zjstudent@126.com

Hong Liu
Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, College of Information Science and Engineering
Shandong Normal University
Jinan, China

***Abstract*—A novel Quantum-behaved Particle Swarm Optimization algorithm with probability (P-QPSO) is introduced to improve the global convergence property of QPSO. In the proposed algorithm, all the particles keep the original evolution with large probability, and do not update the position of particles with small probability, and re-initialize the position of particles with small probability. Seven benchmark functions are used to test the performance of P-QPSO. The results of experiment show that the proposed technique can increase diversity of population and converge more rapidly than other evolutionary computation methods.**

Keywords- particle swarm optimization algorithm; quantum-behaved; probability; benchmark function

I. INTRODUCTION

Particle swarm optimization (PSO) , which is used diffusely for its simple concept, is an evolutionary computation technique developed by Dr. Eberhart and Dr. Kennedy in 1995[1]. However, the algorithm cannot converges to the global minimum point with probability one under suitable condition [2]. Jun Sun et al have proposed a global convergence-guaranteed PSO algorithm [3, 4, 5], QPSO algorithm, which is inspired by quantum mechanics. It has been shown that QPSO outperforms PSO on some aspect, such as simple evolution equation, more few control parameters, fast convergence speed, simple operation and so on [6, 7]. However QPSO as well as PSO, confront the problem of premature convergence in multi-modal functions with higher dimensions, which results in great performance loss sub-optimal solutions. In order to further improve QPSO's performance on complex optimization problems, we present a novel algorithm which introduces probability to QPSO (P-QPSO).

In order to test the improved QPSO algorithm (P-QPSO) on multidimensional optimization, nonlinear function minimization is considered in this paper, where the proposed algorithm is applied. Seven benchmark nonlinear functions minimization, including the unimodal and multimodal problems, are selected to optimize. The benchmark functions can efficiently validate P-QPSO's performance regarding convergence to the global solution. This way, we can truly evaluate the contribution and the significance of the improved QPSO particularly over multimodal optimization problems in high dimensions.

The rest structure of this paper is as follows. In section 2, a brief introduction of the QPSO is presented. The novel algorithm is described in section 3.Then the experiment results are given in section 4. Finally, the conclusion is put forward in section 5.

II. QUANTUM-BEHAVED PARTICLE SWARM OPTIMIZATION

In PSO, the population with M individuals, which is treated as a particle, is called a swarm X in the D-dimensional space. For a given solution space, each particle represents a possible solution vector. The position vector of particle i at the generation t represented as $x_i(t) = (x_{i1}(t), x_{i2}(t), \cdots, x_{iD}(t))$, and the velocity vector represented as $v_i(t) = (v_{i1}(t), v_{i2}(t), \cdots, v_{iD}(t))$ .The particle moves according to the equations:

$$v_i(t+1) = wv_i(t) + c_1 r_1 (pbest_i - x_i(t)) + c_2 r_2 (gbest - x_i(t)) \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (2)$$

Where $i=1, 2... M$. w is the inertia weight. c_1 and c_2 are called the acceleration coefficients. r_1 and r_2 are random number uniformly distributed in $(0,1)$. The personal best position ($pbest_i$) is the best previous position of particle i. The global best position ($gbest$) is the best particle position among all the particles in swarm X.

In Newton mechanics space, the state of particle is depicted by its position vector and velocity vector. However in quantum mechanics space, the position and the velocity of a particle cannot be determined simultaneously according to *uncertainty principle*. Therefore the quantum state of a particle is depicted by wavefunction. Inspired PSO in quantum mechanics space, QPSO algorithm is proposed. The equations are as follows:

$$p_i = \varphi \times pbest_i + (1-\varphi) \times gbest \quad (3)$$

$$mbest = \left(\frac{\sum_{i=1}^{M} pbest_{i1}}{M}, \frac{\sum_{i=1}^{M} pbest_{i2}}{M}, \cdots, \frac{\sum_{i=1}^{M} pbest_{ij}}{M}\right) \quad (4)$$

$$x_i(t+1) = p_i \pm \beta \left| mbest - x_i(t) \right| * \ln(1/u) \quad (5)$$

Where φ and u are random number uniformly distributed in $(0,1)$. p_i is called local attractor. $mbest$ is mean best position of the population. Parameter β is called the contraction-expansion coefficient. In the process of

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.31

iteration, ± is decided by the random number, when it is bigger than 0.5, minus sign (－) is proposed, others plus sign (＋) is proposed.

III. The Quantum-behaved Particle Swarm Optimization Algorithm with Probability (P-QPSO)

In QPSO, the fast information flow between particles seems to be the reason for clustering of particles. Diversity declines rapidly, leaving the QPSO algorithm with great difficulties of escaping local optima. Consequently, the clustering leads to low diversity with fitness stagnation as an overall result. We proposed a novel method to address it. In our method, all the particles keep the original evolution with large probability C_r, do not update the position of particles with small probability C_s, and re-initialize the position of particles with small probability C_t. The iteration process is described step-by-step below.

Step 1: Initialize the positions of each particle, personal best positions *pbest* and global best position *gbest* .

Step 2: For each particle, update the local attractor p_i with (3) and update *mbest* with (4).

Step 3: If the random number $r \leqslant C_r$, compute the new positions of the particle according to (5). If the random number $C_r < r \leq C_r + C_s$, the new positions of the particle is not updated. Otherwise if the random number $C_r + C_s < r \leq C_r + C_s + C_t$, the new positions of the particle is re-initialized.

Step 4: Evaluate the objective function value of the particle, and compare it with *pbest* and *gbest* . If the current objective function value is better than that of *pbest* and *gbest* , then update *pbest* and *gbest* .

Step 5: Repeat step 2~4, until the stopping criterion is satisfied or reaches the given maximal iteration.

IV. Experiment Results and Discussion

A. Test Function

Seven benchmark functions, which are divided into two groups [7, 8], are selected to test the performance of P-QPSO. The function expression, search ranges and initialization range are listed in Table 1. The first group have two unimodal functions f_1 and f_2. Function f_1: *Sphere* is simple and easy to solve. Function f_2: *Rosenbrok* has a narrow valley in the area from the local optima to the global optimum. It can be difficult to obtain the global optimum. The second group has five multimodal functions f_3- f_7. Function f_3: *Ackley* has many local optima around the narrow global optimum basin. Function f_4: *Griewank* is a nonlinear multimodal functions. Function f_5: *Weierstrass* is only differentiable on a part of points though it is continuous. Function f_6: *Rastrigin* is a very complicated multimodal problem with a large amount of local optima. It is very easy for algorithms to trap into a local optimum during the search process. Hence the population intelligence algorithm being capable of keeping the diversity of the population effectively can yield a better solution. Function f_7: *Schwdfel* is due to its deep local optima far from the global optimum. These functions are all minimization problems with minimum value zero. The global best position of these seven functions is [0,0,…,0] except function f_2 and f_7. The global best position of f_2 is [1,1,…,1], and that of f_7 is [420.96, 420.96,…, 420.96].

B. Parameter Settings and Experimental Environment

The numerical experiments are conducted to compare PSO, QPSO, P-PSO and P-QPSO algorithms on the seven test functions. In order to investigate the performance of algorithms on high dimensional optimization problems, the population size is set to 20 because too much population size can increase the time complexity of optimization problems. The maximum generation is set to 8000 generations for the dimensions 10, 40, 80 and 100.

The algorithms parameters settings are described as follow: For PSO and P-PSO, the acceleration coefficients are set to $c_1 = c_1 = 2$ and the inertia weight ω is decreasing linearly from 0.9 to 0.4. In experiments for QPSO and P-QPSO, the value of β varies from 1.0 to 0.5 linearly over the running of the algorithm. In P-PSO and P-QPSO, the parameter of probability C_r is increasing linearly from 0.8 to 0.95. The parameter of probability C_s is decreasing linearly from 0.05 to 0.03. The parameter of probability C_t is decreasing linearly from 0.15 to 0.02.

All experiments were run 30 independent times on an Intel(R) Core(TM) i5-3470 CPU @3.20GHz 3.20GHz， 4GB RAM computer with the software environment of MATLAB20013a. The mean values and standard deviation of the results are recorded.

C. Experimental Results and Discussions

Table 2 to table 5 presents the mean values and standard deviation of best objective function values for 30 runs of the four algorithms on the seven test functions with *D*=20, 40, 80 and 100 respectively. The best results among the four algorithms are emphasized in bold.

In low dimensional optimization problems, as can be seen from Table 2, we can obtain some analysis results. For first group, on function f_1: *Sphere*, which is a simple symmetry function and easy to obtain the global optimum solution in search space, the QPSO performs better than other algorithms. The result of PSO is better than P-PSO and P-QPSO because P-PSO and P-QPSO may enhance global search ability but increase the search complexity of optimization problems. For the function f_2: *Rosenbrok*, its optimal solution lies in a narrow area that the particles are always apt to escape. The experiment results on the function show the superiority of P-QPSO, whose results are the best among all algorithms. The second group, multimodal functions, the performance of P-PSO is the worst on all functions, and other algorithms have the same results. For function f_5: *Weierstrass,* f_6: *Rastrigin* and f_7: *Schwdfel*, the

P-QPSO generated best results and high precision than other algorithms.

The complexity of function is enhanced with the increasing dimension of function. We record the results of each function with dimension 40, 80 and 100 in Table 3 to Table 5. It is evident that P-QPSO has best performance with dimension 80 and 100 for functions f_1-f_6. QPSO has better results than other algorithm with dimension 40 for function f_1-f_4. However the result of Q-QPSO is the best for function f_5 and f_6. For function f_7, QPSO has the best performance on high dimensional problems. All the results show that P-QPSO has better global search ability than other three algorithms and can escape strongly the local minimum in most high dimensional optimization problems.

Fig.1 shows the convergence process of four algorithms on the seven test functions with population size 20 and dimension 100. The convergence graph of function f1 is illustrated in Fig.1 (a). However too fast convergence speed leads to the unclear graph. The logarithm operation of iterations is adopted to show the convergence graph in focus. Fig.1 (b) is the new convergence graph with logarithm operation of iterations. As can be illustrated from Fig.1 (b) we can see that the convergence of P-QSPO is more obvious in new graph than that of Fig.1 (a). Fig.1(c)-(h) are the convergence graphs of four algorithms for other functions.

TABLE 1. TEST FUNCTIONS

Function	Function Expression	Search Domain Initial Range
Sphere	$f_1(X)=\sum_{i=1}^{n} x_i^2$	[-100,100] [-100,50]
Rosenbrock	$f_2(X)=\sum_{i=1}^{n-1}(100\cdot(x_{i+1}-x_i^2)^2+(x_i-1)^2)$	[-2.048,2.048] [-2.048,2.048]
Ackley	$f_3(X)=e+20-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D}x_i^2}\right)-\exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)\right)$	[-32.768,32.768] [-32.768,16]
Griewank	$f_4(X)=\frac{1}{4000}\sum_{i=1}^{n}x_i^2-\prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}})+1$	[-600,600] [-600,200]
Weierstrass	$f_5(X)=\sum_{i=1}^{D}\left(\sum_{k=0}^{k\max}[a^k\cos(2\pi b^k(x_i+0.5))]\right)-D\sum_{k=0}^{k\max}[a^k\cos(2\pi b^k\cdot 0.5))]$ $a=0.5, b=3, k\max=20$	[-0.5,0.5] [-0.5,0.2]
Rastrigin	$f_6(X)=\sum_{i=1}^{n}(x_i^2-10\cdot\cos(2\pi x_i)+10)$	[-5.12,5.12] [-5.12,2]
Schwefel	$f_7(X)=418.9829n-\sum_{i=1}^{n}x_i\sin(\sqrt{\lvert x_i\rvert})$	[-500,500] [-500,500]

TABLE 2 NUMERICAL RESULTS OF DIMENSION 10 FUNCTIONS

function	PSO Mean (St. Dev.)	QPSO Mean (St. Dev.)	P-PSO Mean (St. Dev.)	P-QPSO Mean (St. Dev.)
f_1	1.564E-181 **(0)**	**4.941E-324** **(0)**	3.94E-5 (3.07E-5)	4.42E-34 (1.11E-33)
f_2	0.9708 (0.8448)	0.3716 (0.4878)	5.8804 **(0.3019)**	**0.2654** (0.4639)
f_3	2.665E-15 **(0)**	**2.309E-15** (1.094E-15)	1.949E-03 (8.596E-04)	2.842E-15 (7.944E-16)
f_4	0.0718 (0.0337)	**0.0349** **(0.0167)**	0.1675 (0.0974)	0.0434 (0.0263)
f_5	**0** **(0)**	**0** **(0)**	0.0266 0.0163	**0** **(0)**
f_6	2.3879 **(1.5615)**	2.2611 (1.5949)	3.7628 (2.0431)	**0.8090** (2.0360)
f_7	1173.4445 (383.6300)	694.4573 **(173.6380)**	1299.3772 (301.5361)	**363.4300** (205.3477)

TABLE 3 NUMERICAL RESULTS OF DIMENSION 40 FUNCTIONS

function	PSO Mean (St. Dev.)	QPSO Mean (St. Dev.)	P-PSO Mean (St. Dev.)	P-QPSO Mean (St. Dev.)
f_1	2.738E-24 (7.222E-24)	**3.825E-37** **(1.381E-36)**	2.13E+01 (1.16E+01)	1.14E-17 (1.73E-17)
f_2	49.2815 (30.4139)	**33.4488** **(1.1790)**	89.9610 (44.4388)	35.6790 (16.4522)
f_3	9.031E-13 (1.560E-12)	**2.576E-14** **(1.178E-14)**	3.107E+00 (6.335E-01)	3.319E-10 (1.239E-10)
f_4	0.0107 **(0.0117)**	**0.0075** (0.0121)	1.2929 (0.1832)	0.0088 (0.0146)
f_5	0.2047 (0.4546)	0.0002 (0.0005)	5.9177 (1.2684)	**0** **(0)**
f_6	60.1969 (13.0521)	33.6854 **(6.2650)**	138.2370 (29.3575)	**29.4593** (9.5400)
f_7	7539.5740 (915.4485)	**4327.6259** **(806.4964)**	7790.5304 (950.1597)	5873.0862 (2892.6589)

TABLE 4 NUMERICAL RESULTS OF DIMENSION 80 FUNCTIONS

function	PSO Mean (St. Dev.)	QPSO Mean (St. Dev.)	P-PSO Mean (St. Dev.)	P-QPSO Mean (St. Dev.)
f_1	3.981E-05 (1.416E-04)	3.700E-09 (7.369E-09)	3.394E+03 (1.416E+03)	**1.215E-09** **(1.055E-09)**
f_2	149.0544 (43.0572)	88.7198 (29.5027)	663.4943 (133.1724)	**75.3139** **(12.5883)**
f_3	1.7203 (0.5345)	**0** **(0)**	10.1405 (1.3533)	**0** **(0)**
f_4	0.0142 (0.0282)	0.0041 **(0.0073)**	28.4086 (7.8540)	**0.0038** (0.0080)
f_5	7.0504 (2.6499)	0.1766 (0.1544)	37.1059 (4.1631)	**0.0542** **(0.0259)**
f_6	216.6060 **(31.8994)**	161.7935 (34.0933)	544.6748 (77.9585)	**147.8676** (36.5467)
f_7	18905.3380 (1772.8580)	**9275.7706** (1529.6894)	19378.5490 **(1457.0408)**	15817.7960 (7636.5824)

TABLE 5 NUMERICAL RESULTS OF DIMENSION 100 FUNCTIONS

function	PSO Mean (St. Dev.)	QPSO Mean (St. Dev.)	P-PSO Mean (St. Dev.)	P-QPSO Mean (St. Dev.)
f_1	2.251E-01 (9.244E-01)	5.732E-05 (8.438E-05)	8.438E-05 (2.000E+03)	**1.243E-07** **(6.091E-08)**
f_2	210.9228 (66.1553)	119.9300 (33.0146)	1920.2460 (343.4815)	**109.2489** **(28.4086)**
f_3	2.7527 (0.9759)	0.0061 (0.0065)	14.2954 (1.8004)	**0** **(0)**
f_4	0.0608 (0.0885)	0.0044 (0.0053)	88.7655 (22.7981)	**0.0017** **(0.0037)**
f_5	11.9897 (3.7185)	0.9024 **(0.4174)**	63.6264 (7.0722)	**0.5815** (0.5655)
f_6	310.8314 (40.4044)	252.2316 (44.8682)	866.6371 (108.6755)	**220.5230** **(36.1050)**
f_7	24664.1650 **(1731.0162)**	**13175.8820** (4854.8901)	26144.0420 (1848.0602)	21107.6930 (10395.6060)

Comparing the convergence performance of four algorithms population size 20 and dimension 100, we can see that the performance of P-QPSO, not only the convergence speed but also the best function value, is best among all four algorithms for function f_2 and f_4. At the initial iterations, the result of P-QPSO is worse than QPSO for function f_1, f_3, f_5 and f_6. However the last results of P-QPSO is better than QPSO.

From the results above in the tables and figures, it can be concluded that the P-QPSO has better global search ability than PSO, QPSO and P-PSO for most high dimensional functions.

V. CONCLUSIONS

In our method, to keep the diversity of swarm at the later period of iteration, all the particles keep the original evolution with large probability, do not update the position of particles with small probability, and re-initialize the position of particles with small probability. It can be extend the search space and converge to optimum solution fleetly. The results of experiments have showed that, contrast to other algorithm, the P-QPSO algorithm performs better on global convergence and has stronger ability to escape from the local optimal solution during the search process especially with the high dimension multimodal functions. However with the increasing complexity of the problem, time consumption when solving multimodal problems is the main deficiency of P-QPSO.

ACKNOWLEDGMENT

This work was supported by *A Project of Shandong Province Higher Educational Science and Technology Program* (J15LN03) and *Combined Open Foundation of Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology*.

REFERENCES

[1] J. Kennedy and R. Eberhart, "Particle Swarm Optimization", Proc. IEEE International Conference on Neural Networks (ICNN 1995), IEEE Press, Nov. -Dec. 1995, pp. 1942-1948, doi: 10.1109/ICNN. 1995. 488968.

[2] F. Van den Bergh, "An Analysis of Particle Swarm Optimizers", Ph.D. thesis, University of Pretoria, South Africa, 2002.

[3] J. Sun, B. Feng, and W. B. Xu, "Particle Swarm Optimization with Particles Having Quantum Behavior", Proc. IEEE Congress on Evolutionary Computation (CEC 2004), IEEE Press, Jun. 2004, pp. 325-331, doi: 10.1109/CEC.2004.1330875.

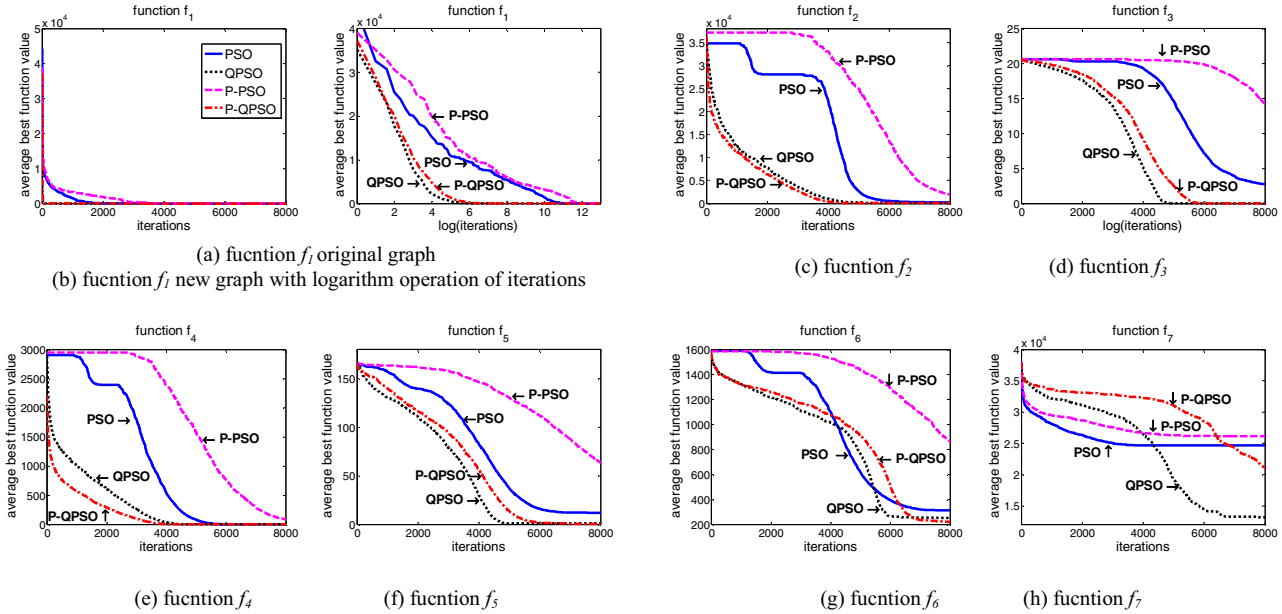
[4] J. Sun, W. Fang, X. J. Wu, V. Palade, and W. B. Xu, "Quantum-behaved Particle Swarm Optimization: Analysis of Individual Particle Behavior and Parameter Selection", Evolutionary Computation, vol. 20, Dec. 2012, pp. 349-393, doi: 10.1162/EVCO_a_ 00049.

[5] W. Fang, J. Sun, Y. R. Ding, X. J. Wu, and W. B. Xu, "A Review of Quantum-behaved Particle Swarm Optimization", IETE Technical Review, vol. 27, Jul. 2010, pp. 336-348, doi: 10.4103/0256-4602. 64601.

[6] J. Sun, X. J. Wu, V. Palade, W. Fang, C. H. Lai, and W. B. Xu, "Convergence Analysis and Improvements of Quantum-behaved Particle Swarm Optimization", Journal of Information Science, vol. 193, Jun. 2012, pp. 81-103, doi: 10.1016/j.ins.2012.01. 005.

[7] J. Sun, "Particle Swarm Optimization with Particles Having Quantum", Ph.D. thesis, Jiangnan University, Wuxi, China, 2009. (in Chinese)

[8] Tu Z G, Yong L., "A robust stochastic genetic algorithm (StGA) for global numerical optimization". IEEE Transactions on Evolutionary Computation, vol. 5, Oct. 2004, pp. 456-470, doi: 10.1109/TEVC. 2004.83125



(a) fucntion f_1 original graph
(b) fucntion f_1 new graph with logarithm operation of iterations

(c) fucntion f_2 (d) fucntion f_3

(e) fucntion f_4 (f) fucntion f_5 (g) fucntion f_6 (h) fucntion f_7

Figure 1 the convergence graph of fucniton f_1-f_7 on four algorithms (20 particles, 100 dimensions)

Solving the Economic Dispatch Problem with Q-Learning Quantum-Behaved Particle Swarm Optimization Method

Xinyi Sheng
Jiangnan University
Wuxi, China
Sheng_xy@hotmail.com

Wenbo Xu
Jiangnan University
Wuxi, China
xwb@jiangnan.edu.cn

Abstract **In this paper, a Q learning quantum-behaved particle swarm optimization (QPSO) method is proposed to solve the economic dispatch (ED) problem in power systems, whose objective is to simultaneously minimize the generation cost rate while satisfying various equality and inequality constraints. The proposed method enhance the global search ability of the algorithm. The feasibility of the QlQPSO method is demonstrated by three different power systems，compared with the GIQPSO, the MeQPSO, and the SlQPSO in terms of the solution quality, robustness and convergence property. The simulation results show that the proposed QlQPSO method is able to obtain higher quality solutions stably and efficiently in the ED problem than other tested optimization algorithm.**

Keywords-component; QlQPSO; QPSO; economic dispatc; global search ability

I. INTRODUCTION

The power economic dispatch (ED) is one of the most important problems in power system operation. To minimize the total generation cost of the generating units is its target while satisfying various constraints of the units and system. Some models see it as a nonlinear optimization problem considering its nonlinear characteristics including discontinuous prohibited zones, unit power limits, ramp rate limits, and cost functions [1].

In traditional ED problems, the cost function of each generator is approximately represented by a quadratic function and the problem is solved by various mathematical programming methods including the lambda-iteration method, the base point and participation factors method, the interior point method, dynamic programming, and the gradient method [1]-[5]. However, none of these traditional approaches can be able to provide an optimal solution, for they are local search techniques and usually get stuck at a local optimum.

In the past decade, a wide variety of heuristic optimization methods such as genetic algorithm (GA) [6]-[8], particle swarm optimization (PSO) [9]-[11] differential evolution (DE) [12], [13], evolutionary programming (EP) [14]-[16], tabu search (TS) [17], neural network (NN) [18], [19], artificial immune (AI) [20], have been applied in solving the ED problem. Bakirtzis, Petridis and Kazarlis presented a GA method and an enhanced GA to solve the ED problem [6]. According to their work, the results obtained are better than dynamic programming (DP) method. Chen et al. developed a lambda-based GA approach for solving the ED problem, and the method is faster and more robust than the well-known lambda-iteration method in large-scale systems [7]. Chiang suggested an improved genetic algorithm with multiplier updating to solve the ED problem with valve-point effects and multiple fuels [8]. Gaing proposed a PSO method for solving the ED problem in power systems, the simulation results showing that the PSO method is indeed capable of obtaining higher quality solutions than GA method in ED problems [9]. Park et al. designed a dynamic search-space reduction strategy to accelerate the optimization process in the PSO method to solve the ED problem [10]. Coelho and Mariani combined the DE method with the generator of chaos sequences and sequential quadratic programming (SQP) technique to optimize the performance of ED problems [13], and their proposed method outperforms other state-of-the-art algorithms in solving load dispatch problems with the valve-point effect.

Recently, inspired by quantum mechanics, Sun et al. proposed a novel variant of the PSO, called quantum-behaved particle swarm optimization (QPSO) algorithm [21]-[23]. The QPSO outperforms the PSO in global search ability and is a promising optimizer for complex problems.

This paper proposes an improved QPSO with differential Q learning medtod, and explores applicability of the QlQPSO in solving the ED problem. The algorithm has better global convergence characteristic, and its feasibility is demonstrated by three power systems respectively, compared with QPSO, GlQPSO, MsQPSO and SlQPSO.

II. PROBLEM DESCRIPTION

The ED problem in power systems is an optimization problem that determines the power output level of online generator that will result in a least cost system state. It is a nonlinear programming one. Practically, while the scheduled combination units at each specific period of operation are listed, the ED planning must perform the optimal generation dispatch among the operating units to satisfy the system load demand, spinning reserve capacity, and practical operation constraints of generating units that include the ramp rate limit and the prohibited operating zones [24].

A. Formulation of the ED Problem

The objective of the classical ED problem is to minimize the total system fuel cost over some appropriate period (one

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.32

hour typically) while satisfying various constraints, and thus the problem can be defined as the following constrained optimization problem:

$$\text{Minimize} \quad F_{cost} = \sum_{j=1}^{N_g} F_j(P_j) \tag{1}$$

subject to

$$\sum_{j=1}^{N_g} P_j = P_D + P_L \tag{2}$$

$$P_j^{\min} < P_j < P_j^{\max} \quad (j=1,\ldots,N_g) \tag{3}$$

where $F_j(P_j)$ is the cost function of the *j*-th generating unit (in \$/h), P_j is the real output of generating unit *j* (in MW), and N_g is the total number of generating units in this power system.

Equality constraint in (2) means that the total system generation includes load demand of the system and the transmission losses. While the total generation cost is being minimized, the total generation $\sum_{j=1}^{N_g} P_j$ should be equal to the total system demand P_D (in MW) plus the transmission network loss P_L (in MW).

Inequality constraint in (3) requires that the generation of each unit should be between its minimum ($P_j^{\min}$) and maximum ($P_j^{\max}$) production limits, which are directly related to the design of the machine.

The cost function of each generating unit is related to the actual power injected to the system, and is typically modeled by a smooth quadratic function as:

$$F_j(P_j) = a_j + b_j P_j + c_j P_j^2 \tag{4}$$

where a_j, b_j and c_j are the cost coefficients of the *j*-th generating unit.

B. System Transmission Losses

The most popular approach for finding an approximate value of the losses is by way of Kron's loss formula as follows, which represents the losses as a function of the output level of the system generating units:

$$P_L = \sum_{j=1}^{N_g}\sum_{k=1}^{N_g} P_j B_{jk} P_k + \sum_{j=1}^{N_g} P_j B_{j0} + B_{00} \tag{5}$$

where B_{jk}, B_{j0}, B_{00} are known as the loss coefficients or B-coefficients. Using the matrix notation, we can express the loss formula as:

$$P_L = P^T[B]P + B_0 P + B_{00} \tag{6}$$

C. Ramp Rate Limits

A number of studies have focused on the economical aspects of the problem under the assumption that unit generation output can be adjusted instantaneously. Even though this assumption simplifies the problem, it does not reflect the actual operating processes of the generating unit.

Practically, the operating range of all on-line units is restricted by their ramp rate limits. According to [7], the inequality constraints due to the ramp limits are given:

1) if generation increases

$$P_j - P_j^0 \le UR_j \tag{7}$$

2) if generation decreases

$$P_j^0 - P_j \le DR_j \tag{8}$$

where P_j^0 (in MW) is the previous output power, UR_j (in MW/h) is the upramp limit of the *j*-th generator, and DR_j (in MW/h) is the downramp limit of the *j*-th generator.

D. Prohibited Operating Zone

Due to steam valve operating or vibration in a shaft bearing, the system contains some operating zone. In the actual power system, the unit loading must avoid the prohibited zones. The feasible operating zones of unit *j* can be described as follow:

$$\begin{aligned} &P_j^{\min} \le P_j \le P_{j,1}^l \\ &P_{j,k-1}^u \le P_j \le P_{j,k}^l, k = 2,3,\ldots,n_j \\ &P_{j,n_j}^u \le P_j \le P_j^{\max} \end{aligned} \tag{9}$$

where $P_{j,k}^l$ and $P_{j,k}^u$ are the lower and upper bound of the *k*-th prohibited zone of unit *j*, and n_j is the number of prohibited zones of unit *j*.

Combining (7), (8), (9) with (1), (2) and (3), the constrained optimization problem is modified as

$$\text{Minimize} \quad F_{cost} = \sum_{j=1}^{N_g} F_j(P_j) \tag{10}$$

subject to

$$\sum_{j=1}^{N_g} P_j = P_D + P_L \tag{11}$$

$$\begin{aligned} &\max(P_j^{\min}, P_j^0 - DR_j) \le P_j \le \min(P_j^{\max}, P_j^0 + UR_j) \\ &P_j^{\min} \le P_j \le P_{j,1}^l \\ &P_{j,k-1}^u \le P_j \le P_{j,k}^l, k = 2,3,\ldots,n_j \\ &P_{j,n_j}^u \le P_j \le P_j^{\max} \end{aligned} \tag{12}$$

III. QLQPSO TO SOLVE ED PROBLEM

A. QlQPSO Method

Q learning is an enhanced learning method which has nothing to do with the problem model of study, through choosing to maximize the accumulated earnings of agent with a discount, and is the optimal strategies from learning to the agent. According to the Q learning method, we regard particle in the group as a proxy in the heart of the QPSO algorithm and expansion shrinkage coefficient selection

strategy as a collection of agency action, you can achieve quantum particle swarm optimization algorithm and the mapping between the Q learning method. We defined $f_p(a)$ as the corresponding fitness function value of a group of the parent individual and $f_o(a)$ as the corresponding individual fitness function value after adopting the parameter selection strategy a, then returns can be defined as an individual $r(a) = f_p(a) - f_o(a)$ immediately. If in this case particles optional action number (parameter selection strategy) is n, particles produced n after an evolutionary descendants, each offspring adopted n strategy of evolution, number of offspring is n^2, after the m times evolution evolved the same time, the offspring has a total of n^m, to determine a parameter selection strategy, we need exponential operation and the amount of calculation is very large and is not applicable. As for this, we simplify the problem, first evolved to produce n new particles; The particle n evoluties again, each individual generates n new particles, the particles generated by calculating each individual income immediately, this set is obtained by using the Boltzmann distribution again to generate new particles are preserved.

$$p(a_i) = \frac{e^{\frac{r(a_i)}{T}}}{\sum_i^n e^{\frac{r(a_i)}{T}}} \quad (13)$$

According to the calculated probability, choose one of the new individual preserved, offspring that are reserved produce offspring in the same way and retain one of the offspring. Through this kind of simplification, after the m times evolution, individual number that a particle cumulated is $n + (m-1) \cdot n^2$ and the individual number that are used to calculate Q value need to retain is $m \times n$, then the polynomial time complexity reduced to Polynomial order.

Put immediately returns to the type (15), and after the reduction, we get

$$Q(a) = f_p(a) - (1-\gamma) \cdot f_o(a) - \gamma \cdot (1-\gamma) \cdot f_o(a^{(1)}) - \cdots - \gamma^m \cdot f_o(a^{(m)}) \quad (14)$$

QlQPSO algorithm design is as follows:

1）Initialization parameter Settings: including group number; expansion shrinkage coefficient values range β; number iterative algorithm; the discount factor γ; calculated Q value required steps m forward looking; randomly generated initial solution $x(i)$, and set up $pbest(i) = x(i)$, and calculate the global optimal value $gbest$.

2）Compute local attractor $p(i)$.

3）Calculate the average optimal value $mbest$

4) Based on QPSO algorithm iterative equation,update the new position of each particle.

5）According to the learning method of Q,choose the optimal parameter strategy β:

a) For each new particles, using n given particle parameters selection strategy to produce n new offspring, and set $t = 1$;

b)Do While $t < m$

Produce n new offspring, select one of the reserves according to the type (13), make $t++$;

c) Calculate the value of each parameter selection strategy Q and choose to maximize parameters selection strategy's value of the corresponding β to the current value β, at the same time give up n-1 other values β;

6) Update location value, and generate a new group

7) Go back to 4) until the end of the loop condition

B. QLQPSO Algorithm Optimization in Power System

To study QlQPSO algorithm validity of scheduling optimization problems in power system, this section uses the QPSO algorithm and its typical scheduling to optimize problems in power system with the optimal solution, and compared with other improved QPSO algorithm optimization. In a power system model, we choose several typical unit system: 6 - the unit system, 15 - the unit system and 40 - the unit system use QPSO and its improved algorithm has gone through simulated test. QlQPSO obtained optimal results in algorithm between a single average optimal value and optimal value calculation.

For large power system of 40 - unit, has the obvious complexity, can reflect the performance of the algorithm better. So here are the simulation results of 40 - unit in the power system. In the power system of 40 - unit , a total of 40 groups of thermal units, when the user needs 85500 mw, including fuel, gas, coal and other units.

By using basic QPSO algorithm, the global optimal value point of interest of QPSO algorithm (GlQPSO), global QPSO algorithm of balance point of interest (MeQPSO), reinforcement learning (QlQPSO) of QPSO algorithm, the learning of QPSO algorithm are tested (SlQPSO), population size is 40, particle dimension and power system units are consistent to 40, evolutionary iteration number is 1000, under the condition of the same run 50 rounds, table1 shows the simulation test results of five kinds of algorithm. Among them, one of the most optimal value was obtained after running 50 rounds of algorithm running, and the average optimal value is the average value of the optimal value in each round after 50 rounds of algorithm running. Figure1 is five kinds of algorithm in optimization of 40 - the evolution of the unit system convergence curves.

IV. CONCLUSION

Three examples in power system, from simple to complex changes, in a 6 - unit system, five kinds of algorithm performance are close; In the 15- unit system, the algorithm shows some differences between GlQPSO and other algorithms; In the 40 - unit system, the gap between algorithms is more obvious, but QlQPSO algorithm shows

optimal performance, and SlQPSO algorithm is still significantly worse than several other algorithms.

[1] IEEE Committee Report, "Present practices in the economic operation of power systems," IEEE Trans Power Appar Syst PAS-90, 1971, pp. 1768–1775.

[2] A. J. Wood and B. F. Wollenbergy, Power Generation, Operation, and Control. New York: Wiley, 1984.

[3] B. H. Chowdhury and S Rahman, "A review of recent advances in economic dispatch," IEEE Trans Power Syst., vol. 5, no. 4, pp. 1248–1259, Apr. 1990.

[4] Z. X. Lianf and J. D. Glover, "A zoom feature for a dynamic programming solution to economic dispatch including transmission losses," IEEE Trans Power Syst., vol. 7, no. 2, pp. 544–550, Feb. 1992.

[5] S. Granville, "Optimal reactive dispatch through interior point methods," IEEE Trans Power Syst., vol. 9, no. 1, pp. 136-146, Feb. 1994.

[6] A. Bakirtzis, V. Petridis, and S. Kazarlis, "Genetic algorithm solution to the economic dispatch problem," Proc. Inst. Elect. Eng.-Gen., Transm. Dist., vol. 141, no. 4, pp. 377-382, July 1994.

[7] P.-H. Chen and H.-C. Chang, "Large-Scale economic dispatch by genetic algorithm," IEEE Trans. Power Syst., vol. 10, no. 4, pp. 1919-1926, Nov., 1995.

[8] C.-L. Chiang, "Improved genetic algorithm for power economic dispatch of units with valve-point effects and multiple fuels," IEEE Trans. Power Syst., Vol. 20, no. 4, pp. 1690-1699, Nov. 2005.

[9] Z.-L. Gaing, "Particle swarm optimization to solving the economic dispatch considering the generator constraints," IEEE Trans. Power Syst., vol. 18, pp.1187-1195, no. 3, Aug. 2003.

[10] J.-B. Park, K.-S. Lee, J.-R. Shin and K. Y. Lee, "A particle swarm optimization for economic dispatch with nonsmooth cost functions," IEEE Trans. Power Syst., vol. 20, no. 1, pp. 34-41, Feb. 2005.

[11] L. S. Coelho, and V. C. Mariani, "Particle swarm approach based on quantum mechanics and harmonic oscillator potential well for economic load dispatch with valve-point effects" Energy Conversion and Management, Volume 49, Issue 11, pp. 3080-3085

[12] R. E. Perez-Guerrero and J. R. Cedeno-Maldonado, "Economic power dispatch with non-smooth cost functions using differential evolution," in Proc. 2005 the 37th Annual North American Power Symposium, pp. 183-190.

[13] L.S. Coelho and V.C. Mariani, "Combing of chaotic differential evolution and quadratic programming for economic dispatch optimization with valve-point effect," IEEE Trans. Power Syst., vol. 21, no.2, pp. 989-996, May 2006.

[14] Y. M. Park, J. R. Won, and J. B. Park, "A new approach to economic load dispatch based on improved evolutionary programming," Eng. Intell. Syst. Elect. Eng. Commun., vol. 6, no. 2, pp. 103-110, June 1998.

[15] H. T. Yang, P. C. Yang, and C. L. Huang, "Evolutionary programming based economic dispatch for units with nonsmooth fuel cost functions," IEEE Trans. Power Syst., vol. 11, no. 1, pp. 112-118, Feb. 1996.

[16] N. Sinha, R. Chakrabarti, and P. K. Chattopadhyay, "Evolutionary programming techniques for economic load dispatch," IEEE Trans. Eovl. Comput., vol. 7, pp. 83-94, Feb. 2003.

[17] W. M. Lin, F. S. Cheng, and M. T. Tsay, "An improved tabu search for economic dispatch with multiple minima," IEEE Trans. Magn., vol. 38, pp. 1037-1040, Mar. 2002.

[18] J. H. park, Y. S. Kim, I. K. Eom, and K. Y. Lee, "Economic load dispatch for piecewise quadratic cost function using Hopfield neural network," IEEE Trans. Power Syst., vol. 8, pp. 1030-1038, Aug. 1993.

[19] K. Y. Lee, A. Sode-Yome, and J. H. Park, "Adaptive Hopfield neural network for economic load dispatch," IEEE Trans. Power Syst., vol. 13, pp. 519-526, May 1998.

[20] T. K. Abdul Rahman, Z. M. Yasin, and W. N. W. Abdullah, "Artificial-immune-based for solving economic dispatch in Power system," in Proc. 2004 Nat. Power and Energy Conf., pp. 31-35.

[21] J. Sun, B. Feng and W.-B. Xu, "Particle swarm optimization with particles having quantum behavior," in Proc. 2004 Congress on Evolutionary Computation, pp.326-331.

[22] J. Sun, W.-B. Xu and B. Feng, "A global search strategy of quantum-behaved particle swarm optimization," in Proc. 2004 IEEE conference on Cybernetics and Intelligent Systems, pp.111-116.

[23] J. Sun, W.-B. Xu, B. Feng, "Adaptive parameter control for quantum-behaved particle swarm optimization on individual level," in Proc. 2005 IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3049-3054.

[24] K. S. Swarup and S. Yamashiro, "Unit commitment solution methodology using genetic algorithm," IEEE Trans. Power Syst., vol. 17, pp. 87-91, Feb. 2002.

TABLE1 RESULTS COMPARISON OF FIVE ALGORITHMS

	QPSO	GlQPSO	MeQPSO	QlQPSO	SlQPSO
Best Value	1.2324e+05	1.2220e+05	1.2384e+05	1.2052e+05	1.2605e+05
Average Best Value	1.2682e+05	1.2635e+05	1.2677e+05	1.2316e+05	1.3611e+05
Standard Deviation	20.4541	40.8464	7.5434	24.9812	9.8639

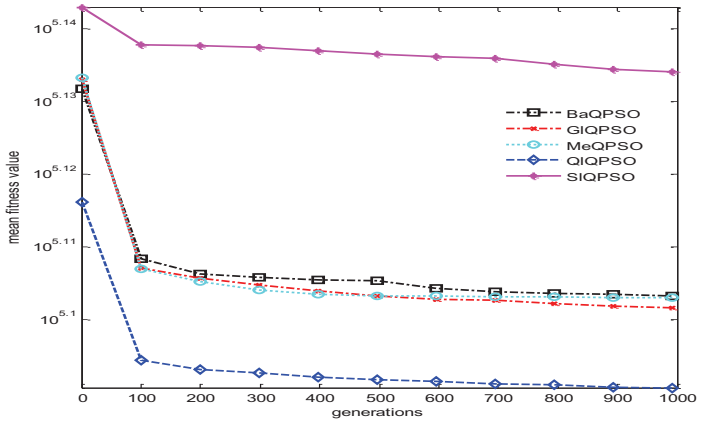


Figure 1. five kinds of algorithm in optimization of 40 - the evolution of the unit system

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Structure Learning Algorithm of DBN Based on Particle Swarm Optimization

Yuansheng Lou
College of Computer and Information
Hohai University
Nanjing, China
wise.lou@163.com

Yuchao Dong*, Huanhuan Ao
College of Computer and Information
Hohai University
Nanjing, China
*Corresponding author: dycwxz@126.com

Abstract: **According to the characteristics dynamic Bayesian network structure, in which the build process, due to the transfer of network nodes is twice the number of variables to be solved, using the traditional method of Bayesian network construction slow efficiency, this article will apply PSO dynamic Bayesian network structure learning, and According to the characteristics Transferred network node in the network is divided into two parts, using stepwise build network has to choose sides proposed DBN structure learning algorithm based on particle swarm optimization. The last benchmark datasets large number of experiments show that the algorithm can improve the efficiency of dynamic Bayesian network structure learning.**

Keywords: dynamic bayesian networks; structure learning; transferred network; particle swarm optimization; Data Mining

I. INTRODUCTION

Probabilistic Bayesian network is a network node in the network represents research questions related variables, the network edges represent interdependencies between variables. In many cases, however, the value of the variable is we change with time and change. Dynamic Bayesian Network (DBN) is a Bayesian network will be extended to represent the time evolution of the process, a new stochastic model having a processing time series data on the original network structure increases the time constraint property formed [1] which has been widely used in speech recognition, highway monitoring, stock market forecast, biological evolution, patient health monitoring and many other aspects.

With the extensive application DBN, many scholars made a lot of dynamic Bayesian network structure learning algorithms, such as genetic algorithm, based on bats swarm algorithm, DBN structure learning algorithm based on quantum particle such as, DBN structure learning efficiency in continuous improvement. Literature [6] gives an algorithm based on ant colony DBN structure learning algorithm, full use of the conditional independence test to limit the search space, learning speed has been relative progress, but it does not meet many needs.

Regarding the characteristics of dynamic Bayesian network structure, particle swarm optimization algorithm is applied to the dynamic Bayesian network structure learning, and the characteristics of the transfer network will be divided into two parts of the network nodes, using stepwise build select network has to the side. The last benchmark datasets large number of experiments show that the algorithm can effectively improve the dynamic Bayesian network structure learning efficiency.

II. DBN AND ITS STRUCTURE LEARNING

Bayesian network is a directed acyclic graph theoretical model to describe the relationship of probability, it provides a representation of causal information. The new dynamic stochastic model based on static Bayesian networks Bayesian networks, based on the original network structure combined with the time information, and the formation of having a processing capability of time-series data.

DBN model state change can be viewed as several frames of animation, the content of each frame is contained in the present time DBN state. DBN in such frames often referred to as time slices, each time slice contains a random set of variables, in order to simplify the problem of enlightenment, it is generally assumed that each time slice is observed random variables belonging to the same subset of variables[2]. Assume that $X=\{X_1, X_2,\ldots,X_n\}$ is a collection of time-varying random variables, X_i [t] represents a variable X_i at time point t corresponding random variables, X[t] denote the set of variables X at time t point corresponding random variables. BDN can be defined as $(B_0, B_{\rightarrow})$, composed by the two Bayesian network. B0 specified initial state probability p(x[0]) of DBN, $B_{\rightarrow}$ specifies transition probability p(x[t]|x[t-1]) which is the probability of state variable set from the time point t-1 to time point t. Because time is not reversible, so there can be a time slice $B_{\rightarrow}$ t to the edge of the time slice t-1. In summary we can see, for a given dynamic Bayesian network B = $(B_0, B_{\rightarrow})$ in X[0], X[1], X[2],..., X[T] probability distribution can be expressed as :

$$p_B(x[0],\ldots,x[T]) = \begin{cases} p_{B_0}(x[0]), \mathrm{t} = 0 \\ p_{B_{\rightarrow}}(x[t]|x[t-1]), 0 \le t \le T \end{cases} \quad (1)$$

By a dynamic Bayesian network B = $(B_0, B_{\rightarrow})$ found that DBN structure learning can be divided into two parts - an initial network B_0 and transfer network $B_{\rightarrow}$, where B_0 is a simple static Bayesian network, $B_{\rightarrow}$ is time slice (t-1, t) changes the state of B is the number of nodes to a static B_0 twice Bayesian network. DBN structure learning can be summarized as follows: for a given set of data samples C, in the definition of an F rating criteria as the basis, to find an optimal network topology structure that search process, namely DBN structure learning. One of the most common rating standards BIC, BDe, K2 and MDL etc [3]. Figure 1 is

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.33

a configuration diagram of DBN contains three nodes, (a) (b), respectively, and the initial network B_0 transfer network $B_{\rightarrow}$, Figure 2 is a corresponding example in three time slice.

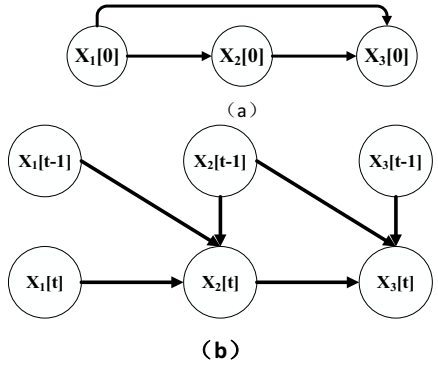


Figure 1 Containing variables X1, X2, X3 of DBN structure graph

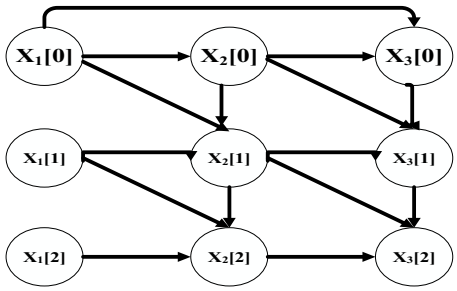


Figure 2 Corresponding instances of three time slices

DBN structure learning should note the following three points:

a) Cross-connection time slice must be directed acyclic graph (DAG);

b) Cross-connection time slice once given, DBN learning process focused on the feature extraction;

c) Learn to connect across time slice is equivalent to choose a variable, and for each node in the network at time t must correspond to select the parent node at time t-1.

III. Structure Learning Algorithm of DBN Based on Particle Swarm Optimization

DBN structure includes an initial network B_0 and transfer network $B_{\rightarrow}$ two parts. B_0 is a simple static Bayesian networks, which already has a lot of classic sophisticated algorithms to build it, and the construct procedures of B_0 described in this article is not expanded. For instance contains n variables, the transfer network $B_{\rightarrow}$ will contain 2n nodes, so in terms of efficiency the request would be higher. We build the network structure in this paper focused on the transfer network, of course this algorithm can also be used to establish the initial network.

A. learning the DBN structure based on PSO

Idea of the Particle Swarm Optimization (PSO) algorithm is to find local optima pBest and global optima gBest in the iteration, and then continue to update the position of the particle through these two extremes. pBest optimal solution is found by a single particle, gBest optimal solution is currently found by entire particle swarm. After Find pBest and gBest particles update their position according to their speed and following two formulas:

$$V_{k+1} = \omega V_k + c_1 r_1(\text{pBest} - X_k) + c_2 r_2(\text{gBest} - X_k) \quad (2)$$

$$X_{k+1} = X_k + V_{k+1} \quad (3)$$

In which k is the number of iterations, and V_k is the current speed of the particle, and X_k is the current position of the particle, and $r_{1,}r_2$ is a random number between 0 to 1, and $c_{1,}c_2$ is learning factor, and ω is the inertia weight used to control the speed of the previous iteration impact on the current speed[4].

1) Position and velocity of the particle representation

Transfer Network $B_{\rightarrow}$ is a directed acyclic graph, defines the location of the particle is the definition of a directed acyclic graph. A use of a matrix to represent directed acyclic graph G, A the elements a_{ij} value 0 and 1, when the node i is parent of node j, a_{ij} value 1, and otherwise value 0.

For the Figure 1 (b), in the order of nodes x_1 [t-1], x_2 [t-1], x_3 [t-1], x_1 [t], x_2 [t], x_3 [t], the matrix A can expressed as:

$$\begin{vmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Similar position is defined, the speed can also be represented as a matrix V, when you add a node i to j directed at the edge, v_{ij} value 1; to delete a node i to j directed at the edge, v_{ij} value - 1; node i to j not delete when it does not increase, v_{ij} value is zero.

2) Position and velocity of the particle updating

Particle position and velocity using equation (2) (3) to be updated before using these formulas need to define subtraction plus operating position P and two operating positions and speed S.

Define the position and speed plus operation P = A + V, depending on the speed of, for the location of particles represented by A have added or deleted to the graph has directed edges, resulting particle group have to figure after the change, the converted directed figure shows a matrix form, that is, the matrix a and matrix V added the resulting matrix P as updated position of the particles.

Set A, B, respectively, after the seed of two positions, define the position of the subtraction S = BA, then when i point j edge node exists A, while B does not exist, sij value 1; if A, edge node i to point j does not exist, while in the presence of B, sij value -1; A and B are the same, then sij equal to 0.

3) Scoring function

MDL metric derived from information theory coding theory, if we want to store sample D in some media, in order to save storage space, naturally you want to store it in a compressed version. And, in order to restore D, you must store the compressed model. Therefore, the total sample description length is defined as D, compressed version of the

total length of the compressed length of the model is described and, MDL principle that the best model is the total length of the shortest description of the model. MDL principle for Bayesian network learning, is to find the network structure with a minimum score of MDL, the network structure describes the joint probability distribution of the sample reflect [5]. According to literature [6] that, MDL score can break down all the nodes in the network of local structures (node associated only with their parents) is the sum of MDL score.

DBN structure contains the initial network B_0 and transfer network $B_{\rightarrow}$, therefore this paper MDL scoring function to find the optimal structure of DBN, DBN of MDL score can be converted to B_0 and $B_{\rightarrow}$ two-part MDL score sum.

B. Stepwise Add the Directed Edge of DBN Structure

When the construction of transfer network $B_{\rightarrow}$, initial graph G only contains nodes. Stepwise select the feasible solution of variables in build process, adding the directed edge for graph G by one increment. we assume that graph G contains n nodes, therefore the number of elected parent nodes of any node is n-1, and the number of elements in the set consisting of parent nodes which is elected by a node in graph G is 2 ^ (n-1). The idea of stepwise adding the directed edge of DBN structure is: Through the analysis of DBN network structure in the previous section, we know that transfer network B can be divided into two time slices t and t-1 and the number of nodes in each time slice is the same. Because time is irreversible, it is impossible to have edge from time slice t to time slice t-1. We divided the set of edge in graph G into two parts E_0 and E_1, while E_0 composed by the directed edges from time slice t-1 to time slice t and E_1 composed by the directed edges among the various nodes in time slice t. When constructing $B_{\rightarrow}$ process, select the set of the parent node for a node in two steps, first graph G will add E_0, the number of candidate parent elements in the set of the current node is 2 ^ (n / 2) then add E_1 to G, increase the number of candidate parent elements in a set of 2 ^ (n / 2-1). Obviously 2 ^ (n-1)> 2 ^ (n / 2) + 2 ^ (n / 2-1), especially when the value of n greater when using stepwise selection to build the network structure to edge ideas $B_{\rightarrow}$ efficiency will be faster.

In the previous section, Figure 1 (b) is a block diagram of DBN contains three nodes of the transfer network $B_{\rightarrow}$, that X_3 [t] increases have added to the side, for example, when using the original method is that its parent node candidate size of the collection of $2^5 = 32$; using stepwise selection to the edge of the network structure thought to be elected its parent node size of the collection of $2^5 + 2^2 = 12$which is much less than 32. Thus, the use of sub-step to select the network side of the ideological structure can greatly reduce the search space, thereby greatly increasing the efficiency of network construction DBN.

C. Realization of The Algorithm

For the set contains n variables, construct the DBN structure B, wherein the transfer network $B_{\rightarrow}$ node contains some (1, 2, ... I, ..., 2n), the proposed algorithm is realized as follows table I:

TABLE I. THE PROPOSED ALGORITHM

The proposed algorithm:

Begin
Initialize position V0,X0,r1,r2,c1,c2, ω;
for t = 0 to k
  for i=1 to n
    for j=n+1 to 2n
      $V_{t+1} = \omega V_t + c_1 r_1(\text{pBest} - X_t) + c_2 r_2(\text{gBest} - X_t)$;
      $X_{t+1} = X_t + V_{t+1}$;
  for i=n+1 to 2n
    for j=i to 2n
      $V_{t+1} = \omega V_t + c_1 r_1(\text{pBest} - X_t) + c_2 r_2(\text{gBest} - X_t)$;
      $X_{t+1} = X_t + V_{t+1}$;
  if($\text{score}_{\text{MDL}}(G_j^{t+1})$< $\text{score}_{\text{MDL}}$(pBest))
    pBest = G_j^{t+1};
  if($\text{score}_{\text{MDL}}(G_j^{t+1})$< $\text{score}_{\text{MDL}}$(gBest))
    gBest = G_j^{t+1};
return gBest;
end

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To test the performance of the proposed algorithm, the use of international standards DBN dataset Alarm (containing 37 node, 46 edge), using Bayes Net Toolbox for Matlab generate sample sizes respectively 2000, 4000, 6000, 8000 four data sets. Experimental runtime environment: Windows7, MATLAB 7.0, Intel core i5, 4GB memory.

We will analysis the comparison of accuracy and the time-consuming respectively in the algorithm proposed in this paper and the ACO-DBN-2S algorithm proposed in literature [6]. To make a fair comparison of the two algorithms are iterative terminated after 500 times. In each data set independently using two algorithms run 10 times, compared to the respective Alarm structure, compared to the average number of correct edge(ANCE) and the average of time consuming(ATC). Results are shown in table II:

TABLE II. COMPARISON OF AVERAGES AND TIME-CONSUMING ABOUT CORRECT EDGE IN TWO ALGORITHMS

Sample Size	The proposed algorithm		ACO-DBN-2S	
	ANCE	ATC (s)	ANCE	ATC (s)
2000	43.5	7.4	38.2	8.8
4000	44.2	8.2	42.8	9.2
6000	44.7	8.5	39.3	9.5
8000	44.9	9.0	40.5	10.3

By table II: In the test sample size for each data set, the average number of correct edges in the proposed algorithm are slightly higher than the ACO-DBN-2S algorithm, and the time consumption is relatively small; when the sample size increases, the proposed algorithm and ACO-DBN-2S algorithm in terms of time consumed is increasing. The average number of correct edges in the proposed algorithm has increasing trend, while ACO-DBN-2S algorithm has not.

This shows that the sample size is larger more significantly the advantage of the proposed algorithm.

V. CONCLUSION

In this paper, particle swarm optimization algorithm is applied to DBN structure learning, for the transfer network will double the number of network nodes, using idea of stepwise adding the directed edge of DBN structure, then proposed a new structure learning algorithm of DBN based on particle swarm optimization. Through large number experiments and compared with other algorithms, the efficiency and time-consuming construct a clear advantage. In future studies, the proposed algorithm should be applied to different areas, such as hydrological forecasting, so apply their knowledge.

ACKNOWLEDGMENT

The research of this paper is partially supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under Grant No. 2013BAB06B04, No. 2013BAB05B00 and No. 2013BAB05B01.

REFERENCES

[1] Chen GuSheng. Based on battlefield information to predict and assess the dynamic Bayesian network [D]. Nanjing University, 2013.

[2] Li Shuo Hao, Zhang Jun. reviewed Bayesian network structure learning [J] Computer Application Research, 2015, 32 (3): 641-646.

[3] Trabelsi G, Leray P, Ayed M B, et al. Dynamic MMHC: A Local Search Algorithm for Dynamic Bayesian Network Structure Learning[M]// Advances in Intelligent Data Analysis XII. Springer Berlin Heidelberg, 2013.

[4] Liu Yi. Improvement and application of the algorithm [D] PSO Xi'an University of Electronic Science and Technology, 2012.

[5] Trabelsi G, Leray P, Ben Ayed M, et al. Benchmarking dynamic Bayesian network structure learning algorithms[C]// Modeling, Simulation and Applied Optimization (ICMSAO), 2013 5th International Conference on. IEEE, 2013:1 - 6.

[6] Hu Renbing, Ji junZhong, Zhang hongXun, etc. ACO based DBN transfer network structure learning algorithm [J]. Computer Engineering, 2009, 35 (22): 191-193.

Study of the Influence of Cross-Border Electronic Commerce on Chongqing's Economic Growth

[1]Qi Wei , [2]Lele Wang
School of Economy and Management, Lanzhou University of Technology, P.R.China, 730050
E-mail:weiq@lut.cn, wanglele1003@163.com

***Abstract*—As a branch of electronic commerce, cross-border electronic commerce is considered as a new commodity transaction pattern. Chongqing, the youngest municipality, possesses excellent infrastructure and policies that support the development of cross-border electronic commerce. An econometric model is established in this paper, based on data related to cross-border electronic commerce that occurred in Chongqing. Through this model, this paper tries to analyze the relationship between Chongqing's economic growth and cross-border electronic commerce, furthermore, testifies that cross-border electronic commerce could promote Chongqing's economic growth on an empirical basis. Finally, based on the results of this study, suggestions are given to promote the development of cross-border electronic commerce in Chongqing.**

Keywords-cross-border electronic commerce; Chongqing, economic growth; model

I. INTRODUCTION

Under the trend of rapid development of electronic commerce globalization, cross-border electronic commerce occupies a more and more important position and become new engine of economic growth. According to IResearch's data,China's import and export volume of cross-border electricity trading amounted 2.3 trillion yuan, up 31.5 percent, accounting for 9.5% [1]of the total import and export volume in 2012. IResearch predicted that cross-border electronic commerce's proportion in China's foreign trade would continue to expand, moreover, the proportion would reach 19.0% and the volume of transaction would reach to 6.5 trillion yuan in 2016. Since March 2012, the Chinese government has attached great importance to the development of cross-border electronic commerce and introduced many documents, such as the "A Number of Views on the Use of Electronic Commerce Platform to Carry out Foreign Trade", "A Number of Opinions of the State Council on the Promotion about Steady Growth and Structure Adjusting of Import and Export", "Notice on the Implementation of the Relevant Policy Advice to Support Cross-border Electronic Commerce Retail Outlet". At the same time, Customs Administration set five cities, Shanghai, Chongqing, Hangzhou, Ningbo and Zhengzhou, as the pilot cities of cross-border electronic commerce in December, 2012.

As the first pilot city, in the early 2008, Chongqing issued a special document "Guidance on Accelerating the Development of the Internet Industry"; pointed out electronic commerce as one of the main areas of development clearly ; and focused on the development of B2B, B2C, network marketing, service of electronic. In October, 2013, with the approval of the General Administration of Customs of the People's Republic of China, Chongqing became the only whole business pilot city of cross-border electronic commerce service, which has four modes including general import, bonded imports, general export and free trade export. In March 7, 2014, Chongqing's international electronic commerce association was formally founded. In June 17, 2014, the business service platform of cross-border trade was founded , which means cross-border electronic commerce of Chongqing started operation. In term of national policies and arrangement, Chongqing has been a pioneer of cross-border electronic commerce in China.

In china, the study of cross-border electronic commerce mainly has focused on four aspects: the development of cross-border electronic commerce in China (E Libin, 2014; Ren Zhixin, 2014), international settlements and logistics facilities of cross-border electronic commerce in China (Cao Shuyan etc, 2013; Huang Ping, 2012), tax jurisdiction of cross-border electronic commerce (Zhang Minwei, 2009; Zhang Xiying, 2011), construction area of cross-border electronic commerce (Tan Bei, 2011). Study on cross-border electronic commerce shows that it makes great difference in realizing the transformation of the foreign trade and promoting economic growth in China. However, there is still short of study on the relationship between cross-border electronic commerce and economic growth. In conclusion, according to the research results, this paper aims to study the cross-border electronic commerce's impact on Chongqing's economic growth from empirical perspective, making some suggestions on the construction of Chongqing cross-border electronic commerce.

II. THE CURRENT SITUATION OF CROSS-BORDER ELECTRONIC COMMERCE IN CHONGQING

A. The definition and characteristics of cross-border electronic commerce

Cross-border electronic commerce is a trading activity based on internet. It has three characteristics. Firstly, it is unlimited, which means trade can happen whenever and wherever possible by cross-border electronic trading platform. Secondly, it is dependent, which means the entire transaction process depends on the Internet. Thirdly, it is effective, which means trading time can be shortened and transaction cost can be saved by cross-border electronic commerce.

B. The development of Chongqing Cross-border Electronic Commerce

Since 2000, the cross-border electronic commerce of Chongqing has been developing rapidly. This paper analyzes

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.34

the development of cross-border electronic commerce in Chongqing, from perspectives of the information flow, logistics and capital flow perspective.

From the perspective of information flow, quantity and quality of hardware and software facilities has been greatly enhanced in Chongqing. In 2012, Chongqing accounted for 1.71% of national IPV4 addresses, number of sites reached to 31437, accounting for 1.2% of number of national sites. Number of Chongqing sites ranked the first of the whole nation at 9.99% a month in update cycle. Now Chongqing has had many different functions of electronic commerce platform and we can see table I for details. F2C (Factory to the Customer) business platform is the largest cross-border electronic commerce platform in Chongqing. Compared to the traditional international trade, F2C platform can reduce the intermediate links among exporters, importers, wholesalers abroad and foreign retailers. Through F2C platform, the goods can be delivered directly from the factory to overseas customers and intermediate expenditures can be reduced.

From the perspective of logistics, Chongqing is known as a mountain city with inconvenient traffic. Nevertheless, with the development of cross-border electronic commerce, comprehensive transportation system of Chongqing has improved constantly. Eurasia International Railway which is an international route from Chongqing to German Duisburg. Generally, shipping goods needs 38 days from China ports to Hamburg ports, however, cargo transportation only needs 16 days through Eurasia International Railway. According to statistics of the Ministry of Railways and the General Customs Administration, 80% freight volume of New Silk Road economy was completed by Eurasia International Railway in 2014. Chongqing Jiangbei International Airport is the third largest airport in the southwest of China. It takes only 10 minutes from Liang Jiang New Area to Jiangbei International Airport.

From perspective of capital flows, many cross-border trading companies of Chongqing use third party payment, such as PayPal, Alipay, 99Bill, Dunhuang net and so on. Different from the traditional one, cross-border electronic commerce based on Internet. In 2012, Chongqing developed cross-border electronic commerce by three measures: building international settlement center, constructing the new electronic commerce park, focusing on transaction certification. In October, 2012, statistics shown that Chongqing provided certification service for more than 70 international business enterprises. Since 2013, Chongqing has built 10 cross-border electronic commerce industrial park. In the Yuzhong industrial park, clusters are initially formed, which consist of Internet information services, electronic commerce services, creative development services and training services[2].

TABLE I. B2B PLATFORM IN CHONGQING

	Classification	B2B platform in Chongqing
1	Marketing service	www.alibaba.com www.hc360.com www.smestar.com
2	Competitive intelligence service	www.mysteel.com www.chemnet.com
3	Online trading services	www.IZP-F2C.com www.dhgate.com www.315.com
4	The third party payment platform	www.yiji.com www.99bill.com www.dhgate.com
5	Data exchange service for foreign trade enterprises	www.quickfund.com www.cqkjs.com

III. EMPIRICAL ANALYSIS

A. Model design

In this paper, in order to study the relationship between the Chongqing cross-border electronic commerce and economic growth, three factors including GDP, import and export volume and the amount of web sites. Because GDP is the most important indicators for measuring economic growth, it is selected as the explained variable. While cross-border electronic commerce is not in the GDP classification project, regional cross-border electronic commerce data is not published in China, so the number of websites in Chongqing and Chongqing import and export volume is seen as the proxy variable of cross-border electronic commerce. The selection of number of sites is based on the fact that domestic and foreign companies usually have a comprehensive understanding of the intentional company through the website. The website plays an important role in cross-border electronic commerce[3].Selection of Chongqing import and export volume is based on the fact that cross-border electronic commerce regards the network as the medium for the import and export business according to the definition of cross-border electronic commerce. So this paper chooses the number of Chongqing site and volume of import and export as the explanatory variable. Chongqing GDP and the volume of import and export from 2000 to 2013 come from National Bureau of Statistics of the People's Republic of China, and the number of sites in Chongqing comes from the China Internet Network Information Center.

In order to avoid the emergence of heteroscedasticity, using log data of GDP, import and export volume and site sequence, this paper establishes the model as follows:

$$\ln GDP_t = \alpha + \beta_1 \ln EX_t + \beta_2 \ln SUM_t + u_0 \qquad (1)$$

Where $\ln GDP_t$ denotes Chongqing's GDP,$\ln EX_t$ denotes Chongqing import and export volume, $\ln SUM_t$ denotes Chongqing sites number, α denotes Constant, u_0 denotes random error.

B. Stationary test of time series

There may be "spurious regression" in time series data, so it is necessary to test the stationary of time series firstly. ADF test is the common method of stationary test to ensure the white noise characteristics of stochastic disturbance. Three ADF test results are shown in table II. The

probabilities of statistics of three time series are all greater than 10%, and the presence of a unit root hypothesis can not be refused, so three time series are non-stationary. After the first order difference, ADF value of lnGDP and lnEX are less than 5% levels, and the ADF value of lnSUM is less than the critical value 10% level. So three series are stationary after the first order difference, and Integrated of Order One I (1) .

TABLE II. MODEL SAMPLE DATA COLLECTION TABLE

variables	ADF	Inspection methods	The1% critical value	The 5% critical value	The10% critical value	Conclusion
lnGDP	-2.3596	（C,T,1）	-4.9893	-3.8730	-3.3820	Not smooth
Δ lnGDP	-4.4617	（C,T,1）	-5.1152	-3.9271	-3.4104	Smooth**
lnEX	-2.1375	（C,T,1）	-4.9893	-3.8730	-3.3820	Not smooth
ΔlnEX	-4.6413	（C,T,1）	-5.1152	-3.9271	-3.4104	Smooth**
lnSUM	-2.1235	（C,T,1）	-4.9893	-3.8730	-3.3820	Not smooth
ΔSUM	-2.8262	（C,T,1）	-5.1152	-3.9271	-3.4104	Smooth***

*** Denotes a coefficient that is significantly different from zero at 1%

** Denotes a coefficient that is significantly different from zero at 5%

* Denotes a coefficient that is significantly different from zero at 10%

C. Cointegration test

Now that lnGDP, lnEX and lnSUM are the same order integration, by using the method of least square, the regression equation is obtained as equation(2):

$$\ln GDP_t = 6.145498 + 0.440055 \ln EX_t + 0.153543 \ln SUM_t \quad (2)$$

(54.32607) (12.24677) (3.573372)

R^2 =0.975778 $\bar{R}^2$ =0.971375 $SE=\hat{\sigma}$=0.111795

DW =1.834613 F =221.5708

In the 5% level of significance, F value is far greater than the critical value of 3.98 which indicates that import and export volume of Chongqing and the number of sites have significant influence on GDP. This can also be seen from the fact that P value is smaller than 0.05. From the goodness of fit, the adjusted coefficient of determination is 0.975778, which shows the ability of explanation of regression equation is 97.58%. That is, import and export volume of Chongqing, the number of Chongqing website can explain the 97.5% change of GDP, which shows the goodness of fit is perfect. Eviews 6.0 is used to calculate the correlation coefficient and partial correlation coefficient. The results are shown in figure 1. Q statistics shows that the p value of each order lag is greater than 0.05 and that partial correlation coefficient histogram does not exceed the dotted line in the 12 lag order, which shows that the null hypothesis can be rejected and the model does not exist autocorrelation under 5% significance levels.

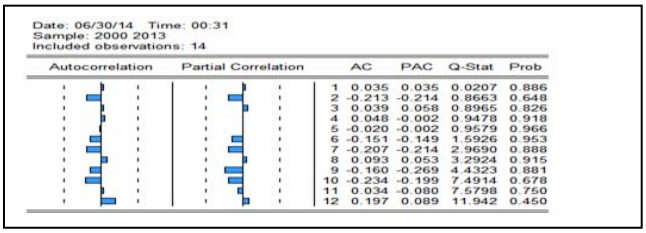


Figure 1. Map of correlation coefficient and partial correlation coefficient

The stationary test results of the residual series is in Table III ,which shows that the residual series reject the null hypothesis at the 5% significance level and residual series of regression equation is stationary. That is, although three variables lnGDP, lnEX and lnSUM,have respective long-term fluctuating rule, there is a long-term stable equilibrium relationship among them.

TABLE III. RESIDUAL TEST RESULTS

variables	ADF	The1% critical value	The 5% critical value	The 10% critical value	Conclusion
Residual	-4.1429	-5.1152	-3.9271	-3.4104	Smooth

D. Granger Test of Causality

Cointegration test results prove that there is a long-term stable equilibrium relationship between Chongqing international electronic commerce development and economic growth. Nevertheless, the causal relationship between economic growth and cross-border electronic commerce needs to be confirmed by Granger test of causality. According to the minimum principle of AIC and SC, the lag order is 4, and the test results are shown in table IV:

TABLE IV. GRAINGER CAUSALITY TEST RESULTS

The primary hypothesis	The observed value	The lag period	F-value	Probability	Result
LnEX not LnGDP of the Granger reasons	10	4	249.952	0.04740	Refuse

LnGDP not LnEX of the Granger reasons	10	4	32.8429	0.13005	Accept
LnSUM not LnGDP of the Gra nger reasons	10	4	3.76805	0.03664	Refuse
LnGDP not LnSUM of the Granger reasons	10	4	12.2498	0.21072	Accept
LnEX not LnSUM of the Granger reasons	10	4	1.23467	0.58098	Accept
LnSUM not LnEX of the Granger reasons	10	4	1.21963	0.58359	Accept

1)The maximum probability that EX is not the Granger reason of GDP less than 0.05, so the null hypothesis can be rejected at the 95% significance level, and EX can be thought as the Granger reason of GDP. The maximum probability that GDP is not the Granger reason of EX, is greater than 0.05,so the null hypothesis can be accepted at the 95% confidence level, and EX is not the Granger reasons of GDP. There is one-way Granger causality between economic growth GDP and EX, which shows that the export growth can promote the economic growth of Chongqing.

2)The maximum probability that SUM is not the Granger reasons of GDP is 0.0366, less than 0.05, That means that along with rapid development of electronic commerce in recent years, Chongqing, as China's inland export processing base and the first zone opening to the outside, has led the rapid economic growth.

IV. THE POLICE SUGGESTION OF IMPROVING CROSS-BORDER ELECTRONIC COMMERCE IN CHONGQING

A Chongqing is one of the cities that have developed rapidly in the cross-border electronic commerce. Results of this empirical study show that there is a stable long-run equilibrium relationship between Chongqing's economic growth and cross-border electronic commerce. Chongqing, as a pilot city, should take the lead in exploring the development model of cross-border electronic commerce and strengthen the leading role of cross-border electronic commerce on economic growth. The existing problems in the construction of Chongqing's electronic commerce include the lag of supervision in Customs clearance, tax, settlement of exchange and drawback, and difficulty in cross-border credit evaluation[4] . In order to promote the development of cross-border electronic commerce in Chongqing, suggestions are put forward as follows.

A. Strengthening the supervision of cross-border electronic commerce

The standardized development of cross-border electronic commerce industry depends on the government supervision. Firstly, Chongqing should set strict market access system to cross-border electronic commerce and establish supervisory organ of cross-border electronic commerce. Secondly, Chongqing should implement register system of cross-border electronic commerce transactions, which must be registered in Customs. With the perfect supervision system, not only the illegal cross-border electronic commerce by some lawbreakers can be prevented, but the related data can be collected to provide data support to the development of Chongqing cross-border electronic commerce.

B. Perfecting assessment of enterprise credit by the third party

Because of serious information asymmetries between both sides of the cross-border electronic commerce, foreign enterprises and Chongqing local enterprises have difficulty in making right credit rating for each other[5]. At first, cross-border electronic commerce enterprises in Chongqing should establish credit filing public websites and foreign and local companies should make credit rating for each other respectively after each transaction on the websites. Then the third party and Chongqing administration of industry and commence should be united to make professional assessments on the local cross-border electronic commerce enterprises of Chongqing and encourage cross-border electronic commerce enterprises to report foreign enterprises' credit to the third party. Meanwhile assessment costs are subsidized by the Chongqing government.

C. Training cross-border electronic commerce talents actively

Because cross-border electronic commerce involves many kinds of products and business process, the cross-border electronic commerce talents are required to master the basic skills including offer, counter-offer, international settlement, international commercial law and business negotiation skills[6]. As a result, the compound talents should be trained for the development of cross-border electronic commerce. Firstly, Chongqing should rely on universities to set up the related majors and cultivate professional talents. Secondly, a mode of cooperation between universities and enterprise should be adopted to encourage universities and enterprises to set up joint training base for cross-border electronic commerce talents and carry out various forms of cooperation. Thirdly, it should be based on all kinds of innovative entrepreneurial projects.

REFERENCES

[1] Ari Consulting Co. Ltd.2012-2013 China Cross-border E-Commerce Report . [EB/OL] . http:/ /report. iresearch. cn/2022. html,2013-09-04.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Xinhua net.com.Electronic commerce enterprises of Yuzhong accounted for half of the country. [EB/OL] http://www.cq.xinhuanet.com/2012-11/15/c_113693296.htm,2012-11-15.

[3] E Libin, Huang Yongwen. "A new way of international trade: The newest research of International Electronic Commerce" ,Journal of Dongbei University of Finance and Economics,2014,(2) : pp.22-31

[4] Meng Xiangming, Tang Qianhui. "Status quo and countermeasure analysis of electronic commerce of cross-border trade in china ",Journal of Shenyang University of Technology (Social Science Edition).2014,(2) : pp.120-125

[5] .Rudolf R. "Sinkovics.electronic commerce and Supply-Chain Management",Journal of Electronic Commer-ce Research.2007,8(4) : pp.221

[6] Nilashi,Mehrbakhsh,Bagherifard,Karamollah.An Application Expert System for Evaluating Effective Factors on Trust in B2C Websites ,Engineering.2011,3,(11),pp.1063-1071

Collaborative filtering Recommendation Algorithm based on MDP model

Wang Xingang
School of Information
Qilu University of Technology
Jinan,China
wxg@qlu.edu.cn

Li Chenghao
School of Information
Qilu University of Technology
Jinan,China
greatlch@126.com

***Abstract*—Collaborative filtering, which makes personalized predictions by learning the historical behaviors of users, is widely used in recommender systems. It makes the prediction and recommend by similarity of users, and it can handle the various work. But the traditional collaborative filtering ignores the connection of users and items. Affect the recommendation's results. To find similar users by measuring the customer relationship between the neighbors can improve the accuracy of prediction user interests'. Then it can improve the accuracy of the recommendation. So collaborative filtering recommendation algorithm based on MDP model is proposed. It can find the connection of users purchase and next purchase. So it can predict users next purchase. Then can recommend items to users. The test results shows the algorithm of this paper have more accuracy.**

Keywords-component; Collaborative filtering; MDP model;

I. INTRODUCTION

There are lots of people want to use the personalized recommendation with the E-commerce's development. So the personalized recommendation algorithms are generated. Collaborative filtering recommendation algorithm is the one which is the most successful. The basic idea of collaborative filtering algorithm is score goals by reference to the project and set high similarity neighbors to predict the target project score, resulting in a final recommendation. With the changes in the structure of the site content and user increased complexity increases ,the systems which bases on collaborative filtering technology are faces some problems such as how to improve the scalability of the algorithm and how to improve the quality of collaborative filtering recommendation algorithm.

Researchers have proposed various solutions. The author propose the two improved similarity measure on paper[1],there are time-based data weight and item similarity-based data weight. The method can reflect the changes in user interests better than others, and it can improve the quality of the recommendation. This paper propose that the traditional emphasis on user similarity may be overstated and there are additional factors having an important role to play in guiding recommendations, and trustworthiness of users must be an important consideration on paper[2]. The author propose that to improve the quality of the recommendation through reduce the dimensionality of databases to solve the problem of sparse data for each project has scores on paper[3]. But this method will cause the data loss. A method is proposed to capture the sequential behaviors of users and items, which can help find the set of neighbors that are most influential to the given users (items) on paper[4]. And it can achieve more accurate rating predictions than the conventional methods. But it's not good for the sparse data. The author propose that collaborative filtering recommendation algorithm based on item clustering on paper[5]. The method improves the real-time performance of recommendation system s by using clustering of user's rating on items. A modified partition clustering is proposed which enhance the accuracy and real-time of recommendation algorithm on paper[6]. In order to improve the recommendation quality of e-commerce system, based on the behavioral characteristics of the user preferences, author propose that recommendation algorithm on feature extraction based on user interests on paper[7]. The method suggests that established a model on feature extraction of user interests according to the dynamic characteristics of network consumer preferences, and designed a corresponding recommendation algorithm. And the other methods such as Bayesian Networks with Hidden Variables[8]、Horting Hatches's graph[9]、association rules[10] have also been used to improve the drawbacks of collaborative filtering technique.

Despite traditional collaborative filtering algorithms can make the appropriate recommendation, but it's often used score information, and ignored timing information and relationship information which is closed. Using the information can in next step enhance the accuracy. Furthermore there are lots of mistakes in the sparse data. So it will affect the quality of recommendation. In this paper we propose the collaborative filtering recommendation algorithm based on MDP model. Firstly, the method using the Markov models to predict the personalized matrix. Second use the tensor decomposition to handle the matrix. Analysis the users interested in the items. Then we will use the collaborative filtering to process the data. So we can receive the better recommended items. Test shows that, we can obtain the high accuracy of the recommendation.

II. TRADITIONAL COLLABORATIVE FILTERING BASED ON ITEMS

Collaborative Filtering based on Items is predicting the score of user to the target object based on user's rating for the similar items. It is based on the assumption that if lots of users have similar scores on the items then they have similar scores on the target items. It's to find the nearest neighbors of the target items by statistical techniques. Then it can use the

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.35

neighbors to predict the scores of the target item.

We denote items set P_u is the user's no score items. For the item $i \in P_u$, the traditional collaborative filtering based on items using under methods. First we should the users set which item i and item j have the similar score. Then using the score to calculate the similarity between item i and item j. The main of similarity measure methods is that Cosine Similarity、Pearson Correlation Coefficient and Adjusted Cosine Similarity[8].Here is the equation.

Cosine Similarity:

$$\mathrm{sin}(i,j) = \cos(\vec{i}\ ,\ \vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 \times \|\vec{j}\|_2} \quad (1)$$

Pearson Correlation Coefficient:

$$\mathrm{sin}(i,j) = \frac{\sum_{u\in U}(R_{ui}-\overline{R_i})(R_{uj}-\overline{R_j})}{\sqrt{\sum_{u\in U}(R_{ui}-\overline{R_i})^2}\sqrt{\sum_{u\in U}(R_{uj}-\overline{R_j})^2}} \quad (2)$$

The users set U is the items i and j who have the same scores. $\overline{R_i}$ is the average scores of item i. User u's rating i is R_{ui}.

Adjusted Cosine Similarity:

$$\mathrm{sin}(i,j) = \frac{\sum_{u\in U}(R_{ui}-\overline{R_i})(R_{uj}-\overline{R_j})}{\sqrt{\sum_{u\in U}(R_{ui}-\overline{R_i})^2}\sqrt{\sum_{u\in U}(R_{uj}-\overline{R_j})^2}} \quad (3)$$

The users set U is the items i and j who have the same scores. $\overline{R_u}$ is the average scores of item u.

Using the equation to calculate the similarity between item i and the others. Put the items who have the higher similarity with item i to become neighbors. Then use the equation to calculate the score of user u to item i. Denote the score is $P_{u,i}$[11] then :

$$P_{u,i} = \frac{\sum_{\mathrm{all}_{\mathrm{sin\,ilar}_{\mathrm{items},N}}}(S_{i,N}\times R_{u,N})}{\sum_{\mathrm{all}_{\mathrm{sin\,ilar}_{\mathrm{items},N}}}(|S_{i,N}|)} \quad (4)$$

The similarity between item I and N is $S_{i,N}$. The user u' highest score of item N is $R_{u,N}$.

III. COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM BASED ON MDP MODEL

A. MDP model based on user's continuous behavior

The model of the paper is that structure the interests of user. Use the user's continuous behavior. Get the user's interesting things and recent preferences. For example:

Table one is the example of continuous behavior. There are four customers who bought items before. The system will recommend based on the last purchase. And the time is not absolute time, but the relative time. For example $t-1$ represent the last purchase.

TABLE I. TABLE TYPE STYLES HISTORY OF USERS PURCHASE

	User 1	*User 2*	*User 3*	*U*ser 4
$t-4$	b, c, d	a, c		
$t-3$	a, d, e	b, e		
$t-2$	b, c	c, d, e	c, d	
$t-1$	b, d, e	a, d	a, b	a, b
t	?	?	?	?

Usually denote the Markov chain which the length is m is $p(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_{t-m} = x_{t-m})$.! (5)

$X_t, \cdots, X_{t-m}$ are random variables. x_{t-m} is their instance. It's not to denote a long chain for all status. So set the m length of Markov chain equals one. After adjusted the denote of Markov chain is $p(B_t^u | B_{t-1}^u)$ . . (6)

The transition probability model of purchase item l to purchase item i is : $a_{u,l,i} = p(i\epsilon B_t^u | l\epsilon B_{t-1}^u)$ (7)

Recommended system gives the user information from the last state. Mining the possibilities of user selects the items. We can denote the average of transition probability model of the last purchase to present purchase.

$$a_{u,l,i} = p(i\epsilon B_t^u | l\epsilon B_{t-1}^u) = \frac{1}{B_{t-1}^u}\sum_{l\epsilon B_{t-1}^u} p(i\epsilon B_t^u | l\epsilon B_{t-1}^u) \quad (8)$$

We can get the Maximum Likelihood $\hat{a}_{u,l,i}$ based on the history of users purchase.

$$\hat{a}_{u,l,i} = \hat{p}(i\epsilon B_t^u | l\epsilon B_{t-1}^u) = \frac{\hat{p}(i\epsilon B_t^u \wedge l\epsilon B_{t-1}^u)}{\hat{p}(l\epsilon B_{t-1}^u)} = \frac{|\{(B_t^u, B_{t-1}^u) : i\epsilon B_t^u \wedge l\epsilon B_{t-1}^u\}|}{|\{(B_t^u, B_{t-1}^u) : l\epsilon B_{t-1}^u\}|} \quad (9)$$

We can obtain the user's transition matrix through the further processing. But these transition matrix are estimated by the small amount of data. And Maximum Likelihood will be under and overfitting on sparse data. The matrix is not reliable in our eyes. So we introduce tensor decomposition to solve this problem. Make the transition matrix into cubes.

TABLE II. USER ONE

	a	b	c	d	e
a	?	1	1	?	?
b	1/2	1/2	?	1	1
c	1/2	1/2	?	1	1
d	1/2	1/2	1/2	1/2	1/2
e	?	1	1	?	?

TABLE III. USER TWO

	a	b	c	d	e
a	?	1	?	?	1
b	?	?	1	1	1
c	1/2	1/2	?	1/2	1/2
d	1	?	?	1	?
e	?	?	1	1	1

TABLE IV. USER THREE

	a	b	c	d	e
a	1	1	?	?	?
b	?	?	?	?	?
c	?	?	?	?	?
d	1	1	?	?	?
e	?	?	?	?	?

TABLE V. USER FOUR

	a	b	c	d	e
a	?	?	?	?	?
b	?	?	?	?	?
c	?	?	?	?	?
d	?	?	?	?	?
e	?	?	?	?	?

B. Tensor decomposition

The model of tensor A is:

$$A = C \times {}_U V^U \times {}_L V^L \times {}_I V^I \quad (10)$$

The core tensor is C. The user's feature matrix is V^U. The previous step selected items' transition matrix is V^L. And the next step selected items' transition matrix is V^I. They have the following structure:

$$R^{k_U,k_L,k_I}, V^U \in R^{|U|\times k_U}, V^L \in R^{|L|\times k_L}, V^I \in R^{|I|\times k_I} \quad (11)$$

k_U, k_L, k_I is analyze dimensions. Denote $k_U = k_L = k_I = k$. Then tensor A is CP(CANDECOMP/PARAFAC). Use CP's pairwise interactions between the three dimensions we can get

$$\hat{a}_{u,l,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \langle v_i^{I,L}, v_l^{L,I} \rangle + \langle v_u^{U,L}, v_l^{L,U} \rangle =$$

$$\sum_{f=1}^{k} v_{u,f}^{U,I} v_{i,f}^{I,U} + \sum_{f=1}^{k} v_{i,f}^{I,L} v_{l,f}^{L,I} + \sum_{f=1}^{k} v_{u,f}^{U,L} v_{l,f}^{L,U} \quad (12)$$

Put (8) and (12) together , we can get the MDP model:

$$\hat{a}_{u,l,i} = \hat{p}(i \epsilon B_t^u | l \epsilon B_{t-1}^u) = \frac{1}{|B_{t-1}^u|} \sum_{l \epsilon B_{t-1}^u} \hat{a}_{u,l,i} = \langle v_u^{U,I}, v_i^{I,U} \rangle +$$

$$\frac{1}{|B_{t-1}^u|} \sum_{l \epsilon B_{t-1}^u} (\langle v_i^{I,L}, v_l^{L,I} \rangle + \langle v_u^{U,L}, v_l^{L,U} \rangle) \quad (13)$$

We can say that $v_u^{U,L}$, $v_l^{L,U}$ are independent of user interaction next purchase items i from (9). If isolate (U, L) ,we can get

$$\hat{a}_{u,l,i}' = \hat{p}(i \epsilon B_t^u | l \epsilon B_{t-1}^u) :=$$

$$\langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \epsilon B_{t-1}^u} (\langle v_i^{I,L}, v_l^{L,I} \rangle) \quad (14)$$

Because (U, L) is independent of i. So to any two items (i, j). $\forall u, t, i, j: \hat{a}_{u,t,i} - \hat{a}_{u,t,j} = \hat{a}_{u,t,i}' - \hat{a}_{u,t,j}'$ (15)

Equation (14) and equation (13) have the same items ranking. This reduces the model's complexity and computation's calculation.

C. Based on Collaborative Filtering Recommendation

We already know the tensor decomposition model. We will introduce how to use collaborative filtering to handle the matrix that we get in the previous section.

We get the users-items transition probability matrix R based on tensor decomposition. Items set is $I = \{i_1, i_2, \cdots, i_M\}$. Users set is $U = \{u_1, u_2, \cdots, u_N\}$. The number of recommended items is p. Target user is u.

Output: Recommended set I_{rec}.

Calculation of users' similarity. Calculation of users' similarity based on the correlation similarity[10] . The result is users' similarity matrix. $R_{sim}(N, N)$. (R_{sim} is N by N square. N denotes users count N. The element's value of it is the main diagonal of axisymmetric distributions. $sim_{k,m} = sim_{m,k}$). The formula of correlation similarity is:

$$sim(u1, u2) = \frac{\sum_{u \in C_{u1,u2}} (R_{u1,i,j} - \overline{R_i})(R_{u2,i,j} - \overline{R_j})}{\sqrt{\sum_{u \in C_{u1}} (R_{u1,i} - \overline{R_i})^2} \sqrt{\sum_{u \in C_{u2}} (R_{u2,j} - \overline{R_j})^2}} \quad (16)$$

The user $u1$ and user $u2$ have the same transition items set is $C_{u1,u2}$. The probability of user bought I in previous step and buy j in next step is $R_{u1,i,j}$ or $R_{u2,i,j}$. The average of user bought i or j is $\overline{R_i}$ or $\overline{R_j}$.

The items set of user u that not buy is $N_k = I - I_k (1 \leq k \leq M)$. The items set is I. The items set of user that bought before is I_k.

Chosen the target user u's closed neighbors set $U = \{u_1, u_2, \cdots, u_p\}$ based on $R_{sim}(N, N)$. Set $u \notin U$ and $sim(u, u_1)$ is the biggest, $sim(u, u_2)$ is the litter than $sim(u, u_1)$, and so on.

Calculation recommended set based on the target user u's closed neighbor p. Use the formula to predict the probability of user buy the item for the every item i that the items set N_k of user u not buy[4].

$$P_{u,i} = \overline{R_u} + \frac{\sum_{u_k \epsilon U} sim(u, u_k)(R_{u_k,i} - \overline{R_{u_k}})}{\sum_{u_k \epsilon U} (|sim(u, u_k)|)} \quad (17)$$

The similarity of user u and its neighbor u_k is $sim(u, u_k)$. The probability of u_k bought i is $R_{u_k,i}$. The average buy probability of user u or user u_k is $\overline{R_u}$ or $\overline{R_{u_k}}$.

Sort of N_k . Make up recommended set $I_{rec} = \{i_1, i_2, \cdots, i_p\}$ using the p number before items to recommend to target user u.

IV. EXPERIMENT EVALUATION

We test the performance of the proposed algorithm in this paper by the experiment. And compared with other algorithm.

A. Environment and data of experiment

The hardware of computer is Interl(R) Core(TM) i5-3470@3.20GHz. The os is windows7. All the program using C implementation.

We used data from our MovieLens recommender system, MovieLens is a web-based research recommender system that debuted in Fall 1977. The site now has over 43000 users who have expressed opinions on 3900+ different movies. It contain 100000 anonymous rating of approximately 1682 movies made by 943 MovieLens users. Each user has rated at least 20 movies. GroupLens provide all datasets while dividing it into five disjoint subsets. Then chose a subsets as test data set and the other four set combined for a base data set. The five pairs of base data set 、test datasets are produced. On this basis , we do the experiment by 5-fold cross-validation. Each time chose a pair of base data set and test data set. The base data set used to do the experiment. The test data set used to do the test of experiment. The average of calculation's error is experiment's error.

Considered data set samples multiplied effect on algorithm performance. We select three data sets by randomly drawn from 200、400 and 600 users' ratings data . Denote as TD200、TD400 and TD600. The table shows the data sets' detail. 80%/20% splits of the data set into training and test data.

TABLE VI. DATA SETS

Data sets	*Users*	*Movies*	*Ratings*
TD200	200	1412	19981
TD400	400	1543	44317
TD600	600	1625	66277

B. Evaluation Metrics

Recommender system research has used several types of measures for evaluating the quality of a recommender system. They can be mainly categorized into two classes: Statistical accuracy metrics evaluate and decision support accuracy metrics evaluate.[12] We used MAE(Mean Absolute Error) as our choice of evaluation metric to report prediction experiments because it is most commonly used and easiest to interpret directly. For each ratings-prediction pair $< p_i, q_i >$ this metric treats the absolute error between them, i.e., $|p_i - q_i|$ equally. The MAE is computed by first summing these absolute errors of the N corresponding ratings-prediction pairs and then computing the average. Formally:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad (18)$$

The lower the MAE, the more accurately the recommendation engine predicts user ratings.

1) Comparison of Similarity Algorithms:

We implemented three different similarity algorithms traditional collaborative filtering, UBI-CF[15] and ours' algorithms tested them on our data sets. The number of neighbors is from 5 to 30.(Interval is 5.) We ran these experiments on our training data and used test set to compute MAE. Fig. 1 shows the experimental results.

The figure 1 shows that CF algorithm uses the highest items' ratings data which is similarity with target item as ungraded items ratings. Because the algorithm uses similarity items, few common score data in the sparse data, finally affect the quality of the recommendation system. The UBI-CF algorithm sets up a similarity model of user background. It can improve the quality of the recommendation by supplement the missing data. But the algorithm doesn't consider the interest of user can change in future. In this paper the we consider the influence of items and users , used MDP model to predict the users transition matrix. So it can put the better neighbors to the user. Then do the recommendation. We get the better result. Make the lower MAE.

2) The performance of the algorithm by dataset's multiplied

Table VI shows, the number of ratings is 19982、44317、66277 in TD200、TD400、TD600. The ration about 1:2:3. This test is on the three datasets. The evaluation metrics is MAE, too. Fig. 2 shows the experimental results.

The Fig. 2 shows that the algorithm has the least MAE on TD600. MAE is lower than lower with the increase data. We consider with the increase data the prediction of MDP model is increase. Make the precise prediction on transition matrix. So the quality of recommendation is better than others. We draw a conclusion from the result. The algorithm of this paper is more effective.

V. CONCLUSIONG

This paper sets up user's transition matrix by timing information of consumption. Handle the sparse data by tensor decomposition. We get a new matrix. Then put the new matrix into collaborative filtering algorithm to do recommend. Accordingly the accuracy of the prediction is improving. After experimental verification the method have better accuracy than the existing recommendation. But because the speed of the MDP model, we spend lots of time on the experiment. So in the future we will improve efficiency of the algorithm.

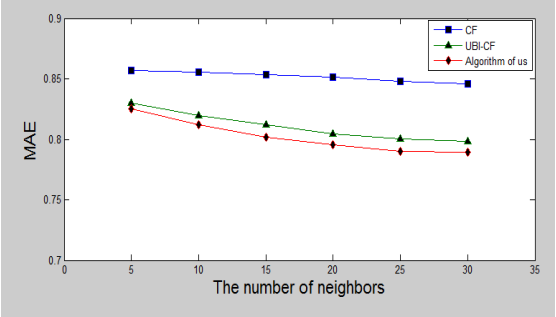


Figure 1: Three algorithms' MAE

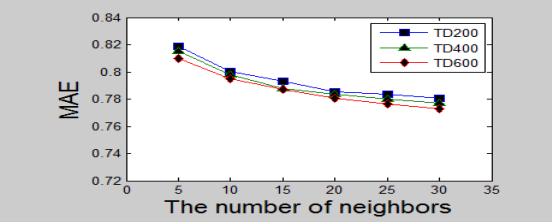


Figure 2: Dataset's effect of the algorithm

REFERENCES

[1] Xing Chunxiao, high Fengrong war Sinan, & Week column. (2007) User Interest Change Collaborative filtering algorithm. Computer Research and Development, 44 (2), 296-301.

[2] Guo Yanhong, Denggui Shi, & Luo rain. (2008), based on trust factor of collaborative filtering algorithm. Computer Engineering, 34 (20), 1-3.

[3] Sarwar, B. (2000). et al., "Application of Dimensionality Reduction in Recommender System-A Case Study. Proc. ACM WebKDD 2000 Web Mining for E-Commerce Workshop.

[4] Sun Guangfu, Wu Yue, Liu Qi, Zhu Chen, & Chen En Red. (2013) based on the temporal behavior of collaborative filtering algorithms. Software University (11), 2721-2733.

[5] Deng Ailin, left cotyledon, & Zhu Yang Yong. (2004). Collaborative filtering recommendation algorithm based on item clustering Small Computer Systems, 25 (9), 1665-1670.

[6] Wu Wang Chen, Xinjun, a brave, & Peng Zhaohui. (2011) based on collaborative filtering and recommendation division improved clustering algorithms Computer Research and Development, 48.

[7] Liu Lin gold, Liu with the deposit, & Bruce Lee. (2011). Recommendation algorithm based on user interest feature extraction. Application of Computer, 28 (5), 1664-1667.

[8] Chickering, D. M., & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. Machine Learning, 29(2-3), 181-212(32).

[9] Aggarwal, C. C., Wolf, J. L., Wu, K. L., & Yu, P. S. (1999). Horting hatches an egg: a new graph-theoretic approach to collaborative filtering.*Proceedings of the Fifth Acm Sigkdd International Conference on Knowledge Discovery & Data Mini*, 201--212.

[10] Sarwar, B., Karypis, G., Konstan, J., & Rield, J. (2000). Analysis of recommendation algorithms for e-commerce. Analysis of Recommendation Algorithms for E-Commerce - ResearchGate, 4(4), 206-207.

[11] Deng Ailin Zhu Yang Yong, & Shi Bole (2003). Based on Item Rating Prediction collaborative filtering algorithms. Software, 14 (9), 1621-1628.

[12] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proc International Conference on the World Wide Web, 4(1), 285--295.

[13] Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized markov chains for next-basket recommendation.. Proceedings of International Conference on World Wide Web Acm, 811-820.

[14] Resnick, P. (1994). Grouplens: an open architecture for collaborative filtering of netnews. Proceedings of the Acm Conference on Computer Supported Cooperative Work, 175--186.

[15] Wu Yifan, & Adrian. (2008) user background information on collaborative filtering algorithm. Computer Applications, 28 (11), 2972-2974.

Research of O2O-oriented service discovery method based on user context

She Qiping [1], *Li Qing* [2], *Deng Juan* [3]
Wuhan University of Science & Technology
City College,
Wuhan, P.R.China, 430083
kuailesqp@163.com[1],
57024241@qq.com[2], 6610773@qq.com[3]

Wu Zhong [4]
Wuhan University of Technology
HuaXia College, Wuhan 430223, China
Wuhan Business University
Wuhan 430056, China
7849800@qq.com[4]

***Abstract*—In the last few years, O2O (Online To Offline) mode become more and more popular, which brings offline service and online web service together, it is important to consider the quality attributes of offline service when recommending service for users online. But in the field of service discovery system, most of the studies focused on function and quality requirements of online service, not on offline service. Furthermore, the existing service discovery methods are generally based on users' function and quality requirements, ignoring the fact that users may have different needs in different contexts. In this paper, we proposed a context-based O2O-ortiented service discovery method, selecting service with function and quality requirements first, and then filtering service with context information, to get the contextual personalized services for users.**

Keywords-user context; online service; offline service; service discovery

Ⅰ. INTRODUCTION

With the increasing prosperity of the online Web service resources, users slide into the plight of "information overload". How to accurately and quickly find out the Web services required by the user and provide effective recommendation has become a great challenge.

At present, research on service discovery is mainly about semantic matching of service function based on service ontology, on this basis, some scholars have expanded the research method and studied service discovery method based on quality of service (QoS) and user context.

However, Current research and Web service application status have the following limits:

(1)User information utilization rate is not high. At present, most of the service discovery methods only consider FQoS (Functional Quality of Service) and QoS (Quality of Service), ignoring the selected service are not the same in different context even if it is the same user[1,2]. Consumer behavior research shows that consumers' context has an important impact on their choice of services.

For example, we must consider user's current location, traffic condition, weather and other "contextual" factors when recommending scenic spot for mobile users[3]. With the development of mobile commerce, there are a growing number of services recommended through the mobile terminal; therefore, the contextual factor is essential for improving the accuracy of service discovery.

(2) The existing web service recommendation method merely focused on online functional attributes(FQoS) and quality attributes(QoS)[4-6]. However, The O2O (Online to Offline) mode is prevalent recently, which brings offline service and online web service together, makes Internet the entry of offline transaction. So the FQoS and QoS of the offline services are also need to be concerned.

To solve these problems, this paper proposed a context-based service discovery method to get the contextual personalized services for users: (1) selecting services with considering FQoS and QoS of both online Web services and offline services; (2) filtering services with context information.

Ⅱ. SERVICE DISCOVERY BASED ON FQOS AND QOS

This section needs to build a ontology meta-model of service by constructing the mapping relationship between FQoS and QoS(shown in figure 1), realize Web service's intelligent recommendation by calculating of the match degree of functional and non-functional attributes.

(1) We analyze functional attributes and non-functional attributes of Web services, and then define the index system structure of and QoS based on the existing theories.

(2) Building the Web service's ontology meta-model, considering concept sets, attribute sets, instance sets and their relationship sets between each other. This paper constructed it from the perspective of online service and offline service, ontology meta-model of service resource is shown in figure 1.

(3) We take a personalized Web service intelligent recommendation algorithm based on users' equilibrium demand[7]. On the basis of Web service's ontology meta-model, we consider influence degree of FQoS and QoS in the process of service recommendation; realize service match between users' needs and providers' items.

Ⅲ SERVICE DISCOVERY BASED ON CONTEXT

This section needs to design the user context classification system and introduce the service filtering method based on context.

(1) Firstly, We defined the dimension structure of context based on the existing theories[8].

(2) Secondly, to get user's browsing behavior, we tracked user's log files which reflected user's interest associated with specific context information. Then we built reasoning rules between service ontology model and user context ontology,

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.36

and discovered service based on user context rules reasoning [9].

(3) Thirdly, we made a further service filtration based on user context clustering when there are no enough reasoning rules for new user at first, since users under similar context are likely to choose the same service, so it is possible to select the most frequently used service among the group which the user is located in, and recommend it to the target user.

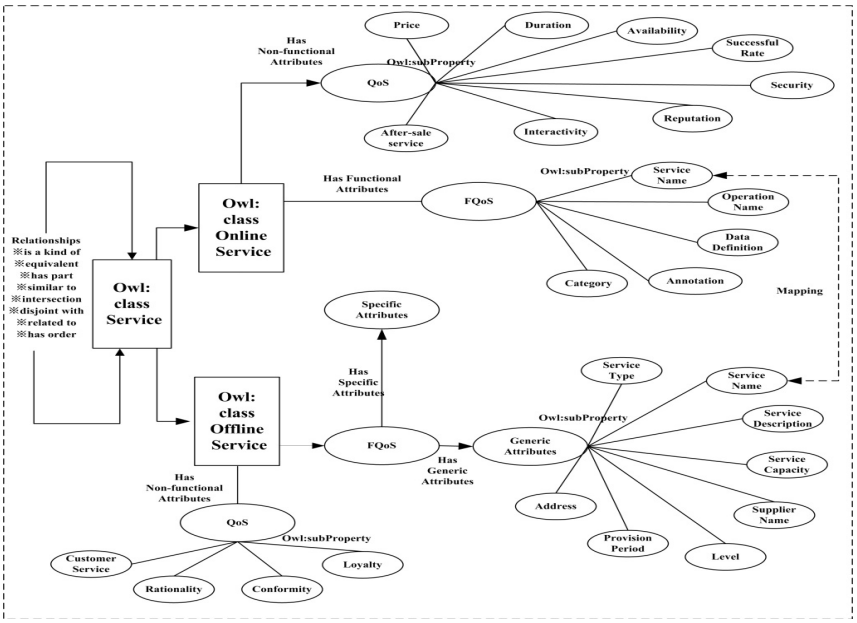


Figure 1. Ontology meta-model of service resource

A. Definition and classification of context information

As defined by DEY, context was "all the information can be used to describe the situation of any entity"[8]. According to the existing research, this section classifies the user context information into three parts: user profile, environment information and platform information.

(1) User profile. It is the description of the user attributes, consisting of user static information and user preference information. Static information includes name, gender, ethnicity, age, occupation, income, educational background, affiliated groups, health, companion, purpose, etc. User preference information refers to a particular lifestyle and personality, for example, some people tend to select a hotel Located in the CBD area, while others favor quiet hotel with elegant environment.

(2) Environment information. It refers to the characteristics of environment around users, consisting of natural environment, social environment and network environment. Natural environment includes location, time, weather, traffic, temperature, season, etc., is the description of the physical information surrounding the user. Social environment includes customs, religion, culture, social trends, legal, social conventions, etc., is the description of user's social environment attributes. Network environment includes network latency, network bandwidth, network security, etc., which describes user‘s. network status

(3) Platform information. It is composed of device information and networking technique. Device information includes desktop, laptop, mobile phone, networking technique refers to all methods that user computing devices access the Internet, such as fixed-line access, 3G networks, 4G networks, wireless LAN, etc..

B. Service discovery based on context reasoning

Service discovery only rely on functional attributes and non-functional attributes, would result in lacking of customization, and ignoring the possible link between the user context and service resources, so it is essential to take a further optimization. This section would use a service discovery method based on context reasoning[9] ,make a further sorting and filtering among the above service discovery results, give priority to return the specific service matching the context. Three important aspects of this method are: user context information, pre-defined rules and rule grades. Context information can be collected and extracted explicitly by asking users for their interest and context information directly or implicitly by tracking users browsing behavior using log files. Reasoning rules between the user context and service resources are divided into three categories:

（1）Filtering rule

Filtering rule means that the service will be filtered out directly when the attribute value of user context and service exceeds the predetermined value.

- v_0 : *value of service attribute* .
- R represents a certain attribute of service.

- "filter S" represents that the service will be filtered out directly, will not appear in the result set returned to the user.

For example: To get plenty time to sightseeing, the night flights will not be selected.

$R(night) \rightarrow filterS$, R=flight time( )

（2）User preference rule

User preference rule means that the service in line with user's preference will be returned to the user.

- $prefereed\ v_s$: $prefered\ value\ of\ attribute\ of\ service$
- v_{si} : $value\ of\ attribute\ of\ service\ i$
- $isChosen(S^n)$ represents n services are selected $(n \in N)$.
- $isPreferred : Q^R$,it represents user prefers some services that the value of attribute meets a certain condition.
- Q represents the attribute of the service meets a certain condition.
- $metBy$:It represents that the preferred value of service attribute can be met by the selected service.

Therefore, user preference rules can be expressed as:

$$isPreferred : Q^R(preferred\ v_s) \rightarrow isChosen(S^n) \wedge \bigwedge_{i=1}^{n} (Q(v_{si}) \wedge$$

$$(isPreferred : Q^R = Q) \wedge (preferred\ v_s\ metBy\ v_{si}))$$

This step will return a set of related services that attributes value meet user preferences.

For example, users want to order tourist attractions tickets three days in advance.

$$isPreferred : Q^R(3\ days) \rightarrow isChosen(S^n) \wedge \bigwedge_{i=1}^{n} (Q(x\ days) \wedge$$

$$(isPreferred : Q^R = Q) \wedge (x\ more\ than\ 3))$$

Q represents how many days in advance can user book attraction tickets.

（3）Optimized selection rule

Optimized selection rule means that when a user context concept is a certain value, the attribute value of the selected service must equal a certain value, or when the attribute value of a certain service meets a certain condition, the service will be selected in priority.

- v_c : $value\ of\ context\ concept\ of\ user$
- v_{si} : $value\ of\ attribute\ of\ the\ i\text{-}th\ service$
- $isChosen(S^n)$ represents n services are selected .
- R represents the value of context concept meets a certain condition.
- Q_j represents the j-th attribute of the i-th service meets a certain condition.

Therefore, optimized selection rule can be expressed as:

$$R(v_c) \rightarrow isChosen(S^n) \wedge \bigwedge_{i=1}^{n} (Q_1(v_{si}) \wedge Q_2(v_{si}) \wedge ... \wedge Q_j(v_{si})),\ n, j \in N$$

For example: When the user with high-income selects hotel reservation service, he/she will prefer large-scale, high-grade and reputable hotel.

$$R(high\_income) \rightarrow isChosen(S^n) \wedge \bigwedge_{i=1}^{n} (Q_1(large) \wedge Q_2(high) \wedge Q_3(great))$$

$R = user\_income, Q_1 = size(), Q_3 = reputation()$。

It means that when a user context attribute is a certain value, it will return a set of related services.

Besides, when the attribute value of some services meets certain condition, those services will be selected in priority.

$$Q(v_{si}) \rightarrow isChosen(S^n) \wedge \bigwedge_{i=1}^{n} (Q_1(v_{si}) \wedge Q_2(v_{si}) \wedge ... \wedge Q_j(v_{si})),\ n, j \in N$$

For example: Flight service will be preferred when there is more than 50% ticket discount,.

$$Q(flight) \rightarrow isChosen(S^n) \wedge \bigwedge_{i=1}^{n} (Q_1(more\ than\ 50\%))$$

Q represents flight services, $Q_1 = discount()$.

Because there may be many rules between user context attributes and service attributes, in order to avoid conflict that may occur in the context reasoning process, we need to specify the order of rank between multiple rules. Because filter rule directly filters out those services which do not meet the needs of context, we consider it has the highest priority. User preference rule represents the subjective view of users when choosing services, we think that the priority of user preference rule is just below filtering rule. Optimized selection rule has the lowest priority.

Let $Rule_1$ denotes filtering rule, $Rule_2$ represents user preference rule, $Rule_3$ represents optimized selection rule, so the priority order of rules can be expressed as: $Rule_1 \succ Rule_2 \succ Rule_3$.

Based on the above context reasoning, we can get a more accurate candidate service set.

C. Service discovery based on user context clustering

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

When user preference information is incomplete, or there are less defined reasoning rules, only a few services can be filtered out by context reasoning, the candidate service set is still too huge for users to select an accurate service. According to the characteristic that users in similar context often make similar choice, this section designs a service selection method based on user context clustering, can find a more appropriate service quickly and efficiently.

（1）Clustering based on user context

Because users in similar context often make similar choice, users can be clustered based on the similarity of user context, thereby forming a number of user groups, when we select service; it is only needed to find the group that the current user is drawn in, further improving the efficiency of service discovery. In this section, we will use multi-particle partitioning method based on fuzzy equivalence relationship [10] to get user context cluster.

For example: We suppose that there are 10 hotel booking services, $S_{hotel} = \{S1, S2, ..., S10\}$, $Ui(i = 1, 2, ..., 9)$ represents old users, U10 is a new user and we need to recommend service for U10. Their context information is

shown in Table 1.

TABLE I. USER CONTEXT INFORMATION

Service	User	user context information			
		Income	*Companions*	*Relationship*	*Purpose*
S1	U1	3	0	0	tour
	U3	6	0	0	tour
S2	U2	4.5	1	friend	tour
	U3	6	0	0	tour
S3	U4	6.5	1	friend	tour
	U5	7	2	colleague	tour
	U6	7	2	family	tour
	U7	8	4	family	tour
S4	U5	7	2	colleague	tour
	U6	7	2	family	tour
S5	U5	7	2	colleague	tour
	U6	7	2	family	tour
	U6	7	2	family	tour
	U7	8	4	family	tour
	U7	8	4	family	tour
S6	U6	7	2	family	tour
	U6	7	2	family	tour
	U7	8	4	family	tour
	U7	8	4	family	tour
S7	U6	7	2	family	tour
	U7	8	4	family	tour
	U8	10	4	colleague	business trip
	U9	15	3	family	tour
S8	U7	8	4	family	tour
	U7	8	4	family	tour
	U8	10	4	colleague	business trip
S9	U7	8	4	family	tour
	U8	10	4	colleague	business trip
S10	U8	10	4	colleague	business trip
	U9	15	3	family	tour
	U10	8	4	family	tour

With the cluster method[10], we got the cluster result, as is shown in Figure 2. We found that U6, U7and U10 are in the same group.

```
>> Untitled
***************threshold value a = 0.985000
group: 1 3
group: 2
group: 4 5 9
group: 6 7 10
group: 8
```

Figure 2. User Cluster

Step three: With the context information and cluster result, we can get the service utilization level of the group, as is shown in table 2, then recommend the Top-5 services (S5，S6，S3，S7，S8）to user.

At this point, we get a more suitable candidate service based on user context clustering, and the most important is that it improves the service discovery efficiency.

TABLE II. TABLE 2. SERVICE UTILIZATION LEVEL

User	service utilization level
U6	S5[2]，S6[2]，S3[1]，S4[1]， S7[1]
U7	S5[2]，S6[2]，S8[2]，S3[1]， S7[1]， S9[1]
U10	

IV. CONCLUSION

Compared with the current service discovery methods which mainly rely on the FQoS and QoS of service, this article conducts a further service selection optimization method based on context reasoning and context clustering, not only gets a more accurate recommendation result, but also improves the time efficiency.

ACKNOWLEDGMENT

This paper is supported by Hubei Provincial Department of Education Science and Technology Research Foundation of 2009CDB032; College fund in Science and Technology Research Expenses of 2013CYYBKY005. Humanity and Social Science Youth foundation of Ministry of Education of China (No.14YJCZH165), the Natural Science Foundation of Hubei Province (2014CFB353)

REFERENCES

[1] Chia-Feng Lin, Ruey-Kai Sheu,Yue-Shan Chang, Shyan-Ming Yuan, “A relaxable service selection algorithm for QoS-based web service composition”, Information and Software Technology, vol. 53, pp. 1370-1381, 2011.

[2] ZHOU Juan,LI Shu-yu, ”A QoS-Based Semantic Web Service Discovery Framework”, Computer Technology and Development, vol. 2, pp. 127-135, 2011.

[3] ZENG Ziming, LI Xin, ”Context - aware Personalized Information Recommendation in Mobile Environment”, Journal of Intelligence, vol. 8, pp. 166-170, 2012.

[4] Ping Wang,KuoMing Chao,ChiChun Lo, “On Optimal Decision for QoS-aware Composite Service Selection”,Expert Systems with Applications, vol. 30, pp. 440-449, 2010.

[5] Vuong Xuan Tran,HideKazu Tsuji,Ryosuke Masuda,“A new QoS Ontology and its QoS-based Ranking Algorithm for Web Services”, Simulation Modelling Practice and Theory, vol.17,pp.1378-1398, 2009.

[6] Min Liu,Weiming Shen,Qi Hao, ”An Weighted Ontology-based Semantic Similarity Algorithm for Web service”, Expert System with Applications, vol. 36, pp. 12480-12490, 2009.

[7] Pingfeng Liu,Hui Zhang,Peilu Zhang, “Method of intelligent recommendation of Web service resources towards users' balanced requirements”, Application Research of Computers, vol. 9, pp. 124-128, 2012.

[8] Anind K Dey, ”Understanding and using context”, Journal of Personal and Ubiquitous Computing, vol. 1, pp. 4-7, 2001.

[9] FENG Zai-Wen, HE Ke-Qing, LI Bing ,GONG Ping, HE Yang-Fan, LIU Wei, ”A Method for Semantic Web Service Discovery Based on Context Inference”, Chinese Journal of Computers, vol. 8, pp. 1355-1363, 2009.

[10] Pingfeng Liu, Yu Wenyan, You Huaijie,”The method of dividing documents into multi-level granules based on fuzzy equivalence relationship”, Journal of The China Society For Scientific and Technical Information, vol. 6, pp. 589-594, 2012.

The Study on the Motivation of T2O E-commerce mode's development

Xu Shuo
The School of Information
Guizhou University of Finance and Economics
Guiyang, China
E-mail: 739972140@qq.com

Yang Yi
The School of Information
Guizhou University of Finance and Economics
Guiyang, China
E-mail: 360544018@qq.com

***Abstract*—T2O is one of the e-commerce business modes which developed from the context of media convergence. The paper explained the essence and motivation of T2O through the research methods of qualitative research, logical inference, case analysis and so on. Firstly, the paper introduced the concept and mechanism of the T2O. Then, it analyzed the differences between the traditional e-commerce modes and T2O, and pointed the uniqueness of T2O. Finally, it launched the motivation of T2O's development from seven perspective, that are the policy-driven, internet thinking, technology-driven, industrial convergence, market demand, business integration and value chain integration.**

Key words-T2O; e-commerce mode; media convergence

I. THE OVERVIEW OF T2O

The triple play of internet, radio and television networks and telecommunications networks not only provides a broad space for the enterprises' development, bring a new round of strategic motivation for the sales industry, but also provides a richer platform for the e-commerce development. T2O is one of the e-commerce modes which developed from the context of media convergence.

T2O (that means TV to Online), is an innovative business model which generated from the context of media convergence. It means that the product sales from TV to online, and refers to the cross-border cooperation of television and Electronic commerce [1]. In other words, T2O achieved the integration of value logical of television and e-commerce, generated a parallel content modules of "video to see + commodity to buy", formed a real-time shopping of "what you see you can buy", and realized the platform's combination, links' communication and users' unity[2].

In fact, T2O's development has long been traceable. The "Fans economy" triggered by the same style with the Stars is one of the most primitive form of T2O. Until the "A Bite of China" which is premiered by CCTV triggered the ratings frenzy in 2012, the e-commerce has caught a ride in this food's program (that is the television program producers cooperated with T-mall and obtained the revenues from T-mall by opening the "tongue train" in it). In this way, T2O created the unique business phenomenon of "orders while watching TV". In August 2014, the variety show "Goddess's New Clothes", which broadcast in Dragon TV, achieved a further cross-border joint of television programs and e-commerce, and opened a new e-commerce model of a watch and buy and Value Instant Conversion successfully.

The mechanism of T2O is that "T" means to promotion and "O" means to sale. If you want "T" transformed into tangible value, it depends on the actual operation and whether the product's quality is good or not [3]. In fact, the basic path of T2O model is that the information spread from TV side to the offline for users to experience through the internet. This model makes the traditional media, electronic business platform, and industry entities into a closed industry value chain, and make the value triggered by the TV side achieved transformation in e-commerce platform by certain platforms or channels.

II. THE DIFFERENCE BETWEEN THE TRADITIONAL E-COMMERCE MODEL AND T2O

In order to understanding easily, I was sorted out the similarities and differences between the traditional e-commerce modes (such as B2B, B2C, C2C, etc.) and T2O e-commerce mode from marketing model, technical means, consumers, merchants, e-business platform, television media, business point of view and the perspective of industry chain, as shown in Table 1.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.37

TABLE 1: COMPARISON OF TRADITIONAL E-COMMERCE MODES AND T2O E-COMMERCE MODE

Compares	**traditional e-commerce models（B2B, B2C, C2C, etc）**	**T2O e-business models**
Marketing Model	Network Marketing	Network Marketing, Video Interactive, Fans Economy
Technology	Based on the large-scale data information exchange in traditional internet	Integration of multiple technologies with triple play, dimensional code scanning, etc
Consumers	A lot of reference information, Screening cumbersome	Multi-directional, Interactive, Multi-dimensional, Free choice of terminals, Active awareness
Merchants	Serious homogeneity competition	More opportunities for promotional display, The promotion effect can be found
E-business Platform	Lack of innovation ability	New development momentum, Improve customer stickiness, Develop new clients
Television Media	Traditional profit model, Single source of income	More attractive to advertisers, Diversify income sources
Enterprise	High-cost, low-interaction	Interactive promotion, content sharing
Industry Chain	Medium single, high transaction costs	Merger competition, mutual benefit and win-win

Source: collecting and sorting

III. THE MOTIVATION OF T2O'S GENERATION

A. Macro Perspective

1) Policy Driven

Policy driven relies on the support of national policy, and plays a supporting and leading role in social economy, improves the contribution of scientific and technological to the national economy, achieves the comprehensive coordination of the economic society and the continuous improvement of the overall national strength. No matter in the national level or organizational level, the support of the policy always plays an irreplaceable role.

The generation of T2O, of course, also relied on the support of national policy. For example, Prime Minister Li who proposed to adjust the industrial structure, support the development of mobile Internet, e-commerce and online financial, and put forward the "Internet +" plan in this year's government work report. These policies promoted the integration of the mobile Internet, cloud computing, networking and traditional industries. The "opinion on developing e-commerce to speed up the new impetus for economic development" which is Printed and distributed by the state council in May 2015, also has the deployment of the innovation and development of e-commerce. In August 2014, the document "Guidelines on promoting the convergence of traditional media and new media" which issued by the central government, proposed that they will create a group mainstream medias of diverse forms, advanced means and new competitiveness, and form a modern communication system of three-dimensional diversity and integrated development. Those various policies promote the innovation and development of e-commerce and traditional media.

2) Internet Thinking

The internet thinking that seeking reform and innovation was showed in the T2O E-commerce mode incisively and vividly. In the book which named "Internet thinking Magic Power" [4], the author expounded nine big Internet thinking, and each of the thinking could be mapped to T2O appropriately[5]. "Users thinking" can catch the bottom of the audience and give them the opportunities to experience. "Simple thinking" can reflect the style of the product and service vividly by the television and provide a simple consumption decision and a convenient purchase process for the consumer. "Extreme thinking" makes the products, services and user's experience to the ultimate to exceed the customer's expectations. "Iterative thinking" means that the

company should focus their thinking on the audience's demand and improve themselves in real time. "Traffic thinking" also helps T2O to aggregate popularity and brings business opportunities to the media industry and e-commerce. Similarly, the "socialization thinking" helps T2O to provide a successful demonstration on social media and crowdsourcing, and changed the whole forms of product's design, production, marketing and sales in one corporate. We can master the user's habits easily by "big data thinking", that is at least 64% audiences playing mobile phones while watching TV, so many of the companies introduced the T2O e-commerce model. It was the core of "platform thinking" that T2O use the existing platform built a more than one subject beneficial ecosystem. Finally, T2O relies on the "cross-border thinking" and brings a different opportunity for its development.

3) Technology Driven

The technology driven refers to the integration and development of technology, including the fusion of the communication terminal, triple play, big data, cloud computing, IOT and 4G etc. With the driving of technology, the mobile shopping has a rapid development in China, and it laid a solid foundation for the production of T2O, providing a broad space for the company's development, and brought a new round of strategic motivation for the sales industry.

4) Industrial convergence

Industrial convergence refers to the boundaries of two or more industries become blurred or even disappear due to the promotion of technology, market, service, management and so on[6]. T2O is the product of industrial convergence which focusing on the internet industry, media industry and e-commerce industry. Its diversified industry consolidation not only provides a breakthrough for the development of media industry, but also provides a new way for the development of the traditional e-commerce [7].

In fact, the performance of the industry convergence is more concentrated in the field of internet. For example, the traditional market plus Internet generated Taobao, traditional department stores plus Internet produced Jingdong, traditional matchmaker plus Internet produced century good marriage, traditional news plus Internet produced ChaiJing's "under the dome" and so on. Therefore, it is not surprised that the internet plus media plus e-commerce produced the T2O e-business mode, and there will be have more products of the industry integration in the future.

B. Micro Perspective

1) Market Demand

Market demand comes from the new features which exhibited by the consumers and audiences' commodity consumption and information acceptance's degree, including scale, audience fragmentation, diversity, timeliness and portability demand. Those characteristics will changed the production and sales form of the traditional goods and information. T2O would achieved the "content means goods" and "real-time shopping" by its innovate ability, and satisfied the internal demand of the e-commerce and media industry's market changes under the background of media convergence. It will becomes a new profit growth point of the media industry and commercial enterprises, and thus to gain the faster growth and higher profits of the enterprise.

In addition, the market demand was also a multidimensional junction points of new e-commerce mode, it makes the website and consumers, institutions and terminals, enterprises and channel agents combined organic according different requirements, then formed a interests and win-win platform. T2O was provided a set of advantages in one emerging e-commerce mode, for example the multimedia integrated promotion, content sharing, volume layout management, social network and internet finance, and it solved the bottleneck of traditional e-commerce model.

2) Business Integration

It is because the effective integration of market that makes the business processes tends to integration. Different forms of product (for example production, distribution, exchange and sales) realized the rapid conversion and transformation through business integration [6]. In addition, business integration includes strategic business integration, structural fusion and so on. Undoubtedly, T2O has extended the value chain of the e-business platform, media industry and the business entities, and integrated various advantages of the media industry, Internet industry and e-commerce.

Taking "Goddess's new clothes" as an example, it was a variety show that broadcast in Dragon TV in August 23, 2014 and the most typical case of T2O. The original mode variety show has been realized the business integration of traditional media and e-commerce. The program brought designers, celebrities and fashion buyers to the show, let the sales platform and user platform to interactive, merged all the processes from design, clothing, shows, auctions to purchases in a same theme, achieved the perfect experience of "what you see you can buy", and was called the landmark of TV shows and e-commerce in internet era[2].

3) Value Chain's Integration

The integration of the value chain makes T2O's operating costs spread to multiple information platforms, businesses groups, sponsors and the general public successfully. As we all known, the value chain of traditional media program is that: contents produced audience--the audience produced the consumer market--the consumer market feedback the special sponsors [2]. So the traditional media could not benefited from the audience directly, the program sponsors could not delivered advertising precisely, and the returns were difficult to estimate either.

T2O has realized the value chain's integration of traditional media and e-commerce. Taking the "Goddess's new clothes" as example either, "Goddess's new clothes" takes the clothing industry as the theme, takes "fashion theme--type design--Clothing modeling--T show--copyright bidding--clothing sales"[8] as the framework, takes the process of the fashion industry as the main line to design the program content, and makes the audiences into real users successfully. It realized the value chain's integration of traditional media and e-commerce, and provided a reference for the development of traditional media, e-commerce and industry entity.

IV. Conclusion

Generally speaking, the T2O's development is not only the change of consumption thinking and service mode, but also a new challenge of the traditional e-business mode and media thinking. It can be said that the integration road is the inevitable trend of industrial development, and is also the fundamental cause of T2O's development. So operators must be changed the inherent thinking timely to realized the convenience of network and the resources' integration, and built a seamless connection between industries, thus creating a new e-commerce mode[9]. Moreover, it is needed to point that, T2O is not the ultimate form in the e-commerce's development. We should believe that, T2O will exceed the existing business mode and found a new field in business area with the online discovered and consumption patterns as well as the consumer tastes and lifestyle changes [10].

References

[1] Cheng Lei, "T2O" business model analysis of "TV + electricity supplier" [J], China Trade, pp. 103, November 2014.

[2] Xu Hongwei, The Mode analysis of the combination of TV and e-commerce [J], Chinese Journal of Radio and Television, pp. 52-53 , January 2015.

[3] Zhang Li, Nan Rui. T2O started, television convergence electricity supplier Sessions [EB / OL], Chinese City of Television Technology Association,

http://info.broadcast.hc360.com/2014/09/250902610898.shtml, 2014.

[4] Zhao Dawei, Internet thinking Magic Power[M], Machinery Industry Press, pp. 2-7, March 2014.

[5] Zhao QianwenThe, transformation of the Internet media industry from the perspective of "the goddess's new clothes"[J], Media observation, pp.10-12, November 2015.

[6] Huang Jin.Media convergence models with contributing factors[M], China Book press, pp. 10-21, January 2011.

[7] Zhang Nan, To exploring the media's convergence mode based on the "Goddess's new clothes" [J], Drama House, pp. 95,October 2014.

[8] Yang Chenying, Goddess' clothes [J] .Chinese Clothing, pp. 89, October 2014.

[9] Su Tao, The analysis of O2O e-commerce business model [J], National Business, pp. 34, January 2012.

[10] Lu Yiqing, Li Chen, O2O business model and prospects research [J], Business Economy, pp. 101, November 2013.

Effects of RMB Exchange Rate Changes on China's Outward FDI

Yu Chao
Management Science and Engineering
Cooperative College
Qingdao Agricultural University
No.700 Changcheng Road Qingdao, China
chaoyuw@163.com

***Abstract*— With the rapid growth and the improvement of international competitiveness of Chinese economy, China's outward FDI has developed quickly, the relationship between RMB exchange rate and outward FDI has attracted more and more attentions in recent years. On the basis of constructing and analyzing the theoretical model, after method of calculation of exchange rate expectation and exchange rate volatility is given, the relationship between RMB exchange rate and China's outward FDI is explored by using the panel data from 2005 to 2013. The result shows that significantly positive correlations exist among the RMB exchange rate level, the exchange rate expectations and China's outward FDI, the RMB exchange rate volatility has significantly negative impact on China's outward FDI, which means the smaller the volatility is, the better promotion of the development of China's outward FDI .**

Keywords- exchange rate; outward FDI; multinational corporation

I. INTRODUCTION

As an effective form of participating in international division and optimizing resources allocation between countries, outward FDI has been accepted by more and more enterprises. The development of multinational corporations has deeply influenced one country's economic progress, industrial structure adjustment and balance of payments. According to the report of UNCTAD, outward FDI has played an important role in promoting national economy, and become the main driving force for global economic integration. Since reform and opening up, depending on labor cost advantage and huge market potential, China has achieved great success in attracting foreign investment, but the development of outward FDI is relatively slow. Until 2001, China's outward FDI begins to show a trend of rapid development because of the increase support for domestic enterprises engaged in overseas investment. As a major force of international investment, nowadays, China has participated in the economic globalization and international division of labor at a higher level, wider scale and deeper depth. The impact of these changes take place on China's outward FDI has attracted researchers' attention for many years.

II. RELATED LITERATURE

A. Change of Exchange Rate Level

Both the theory of relative wealth hypothesis and the theory of comparative cost consider that home currency appreciation will reduce the cost of production abroad and boost investment profits. Gregory and Mccorriston (2005) studied the overseas investment of British enterprises and found that currency appreciation accelerated the development of outward FDI. Some other studies had come to similar conclusions (Sushko, 2007; Udomkerdmongkol and Görg, 2009; Feils and Rahman, 2011).

Different from the above conclusion, some scholars believe that whether multinational corporations invest abroad or not depends on the expected value of returns in the future. In this sense, home currency appreciation will restrict the development of outward FDI, and currency devaluation will promote its development (Görg and Walklin, 2002; Schmidt and Broll, 2009). Other scholars believe that changes of exchange rate have no significant effect on outward FDI or the effect is uncertain (Egger, 2008; Buckley, 2009; Peiming 2013).

B. Volatility of Exchange Rate

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Some scholars believe the volatility of exchange rate increases the uncertainty of cost and profit, it will hinder the development of outward FDI(Barrel 2007). Others think that if the volatility of exchange rate increases, diversification will be a necessary choice, which means domestic enterprises will shift production to lower-cost countries, so volatility of exchange rate will have a positive effect on outward FDI(Liangke, 2012; Gottschalk and Hall, 2008).

Studies about the relationship between RMB exchange rate and outward FDI mainly used time series economic data for many scholars. In fact, the Central Bank of China announced the implementation of managed floating exchange rate system in 2005, based on market supply and demand, regulated through a basket of foreign currencies, the Central Bank of China announced the implementation of managed floating exchange rate system. After that, the exchange rate of RMB has been continuous appreciation against the dollar, and the volatility increased. In this sense, to study the relationship between RMB exchange rate and outward FDI with panel data will be more accurate.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.38

The appreciation of home currency will reduce overseas product costs, and affect future profits, the volatility of exchange rate will increase the uncertainty of future profits, but also promote multinational corporations to choose diversification. The relationship between exchange rate and outward FDI is not the same for different countries and different industries. Even for the same country and same industry, the relationship between them is uncertain in different times. On the basis of constructing and analyzing the theoretical model, panel data will be used to study the relationship between China's outward FDI and its currency exchange rate.

III. THEORETICAL MODEL

Suppose there is a multinational corporation, which will invest in country A and country B, in order to analyze the effect of exchange rate level and volatility on outward FDI, the following assumptions are given. First, the multinational corporation hires locally, which means wages are paid by local currency. Raw material prices, government subsidies and taxes are all expressed in local currency. Second, with the influence of globalization and free competition, valued in home currency, prices of the products produced by the multinational corporation are the same in different countries. The cost of fixed asset investment for multinational corporations of home country can be expressed as $f_i Q_i^2 / 2$, where f_i is investment coefficient. $C_i = C_i(w_i, s_i, P^*, t_i)$ denotes cost of production, where w_i, s_i, P^*, t_i denote wage level, government subsidy, raw material price and tax respectively. So the total investment cost can be expressed as $\frac{f_i Q_i^2}{2} + C_i Q_i$. Assume there is a negative correlation between the effect function and the variance of expected profits of investment and a positive correlation between the effect function and the mathematical expectation of expected profits, the following function is used:

$$U(\pi(\cdot)) = E(\pi(\cdot)) - \varphi Var(\pi(\cdot)) \tag{1}$$

:

The profit function of the multinational corporation can be expressed as the following:

$$\begin{aligned} \pi &= PQ_A S_A - C_A Q_A - \frac{1}{2} f_A Q_A^2 \\ &+ PQ_B S_B - C_B Q_B - \frac{1}{2} f_B Q_B^2 \end{aligned} \tag{2}$$

Where P and Q_i denote price valued in home currency and output produced by the multinational corporation, S_i denotes bilateral exchange rate in indirect quotation between the home country and the host country. From (1) and (2), we can get:

$$\begin{aligned} U(\pi(\cdot)) &= PQ_A E(S_A) - C_A Q_A - \frac{1}{2} f_A Q_A^2 \\ &+ PQ_B E(S_B) - C_B Q_B - \frac{1}{2} f_B Q_B^2 \\ &- \varphi P^2 Q_A^2 Var(S_A) - \varphi P^2 Q_B^2 Var(S_B) \\ &- 2\varphi P^2 Q_A Q_B Cov(S_A, S_B) \end{aligned} \tag{3}$$

Take the partial deviate, we can get

$$Q_A = \frac{PE(S_A) - C_A - 2\varphi P^2 Q_B Cov(S_A, S_B)}{f_A + 2\varphi P^2 Var(S_A)} \tag{4}$$

$$Q_B = \frac{PE(S_B) - C_B - 2\varphi P^2 Q_A Cov(S_A, S_B)}{f_B + 2\varphi P^2 Var(S_B)} \tag{5}$$

$\sigma_A^2 = Var(S_A), \sigma_B^2 = Var(S_B)$ denote exchange rate variances, $\rho = \frac{Cov(S_A, S_B)}{\sigma_A \sigma_B}$ is the correlation coefficient. Take partial derivative to (4), we can get

$$\frac{\partial Q_A}{\partial E_A} = \frac{P(f_B + 2\varphi P^2 \sigma_B^2)}{f_A f_B + 2\varphi P^2 (\sigma_A^2 f_B + \sigma_B^2 f_A) + 4\varphi^2 P^4 (1-\rho^2)\sigma_A^2 \sigma_B^2} \tag{6}$$

Assume the multinational corporation is risk averse, Which means $\varphi > 0$, so $\frac{\partial Q_A}{\partial E_A} > 0$ ($0 < \rho^2 < 1$). The following two conclusions can be got. First, home currency appreciation will promote the development of outward FDI; second, the more the volatility of exchange rate, the smaller value of $\partial Q_i / \partial E_i$ is, which means the same change of exchange rate level will have a smaller effect on the production of the multinational corporation; the less the volatility of exchange rate, the bigger the value of $\partial Q_i / \partial E_i$ is, which means the same change of exchange rate level will have a greater effect on the production of the multinational corporation.

IV. EMPIRICAL STUDY

In order to better study the relationship between China's outward FDI and RMB exchange rate, the following model is given:

$$\begin{aligned} FDIP_{it} &= \alpha_0 + \alpha_1 FDIP_{it-1} + \alpha_2 R_{it} \\ &+ \alpha_3 E(\theta_{it}) + \alpha_4 Vol(\theta_{it}) \\ &+ u_{it} \end{aligned} \tag{7}$$

Where $FDIP = FDI / GDP$ is the ratio of investment flows on host country to GDP of the host country. R_{it} denotes the exchange rate between RMB and the currency of host country in indirect quotation. $E(\theta_{it})$ and $Vol(\theta_{it})$ denote expectation and volatility of exchange rate respectively. Taking account of the availability and continuity of data, panel data of 36 host countries which are the main investment destination of China will be used from 2005 to 2013. The independent variables will be explained as follows:

A. Volatility of Exchange Rate

The effect of exchange rate volatility will be underestimated because of the loss of useful information, to avoid this, some scholars use exchange rate variance as the proxy of volatility, some use standard deviation of 12 consecutive quarters as the proxy of volatility, domestic scholars usually use GARCH model. In fact, As much panel data will be used to study the relationship between outward FDI and exchange rate, the condition of using GARCH model is not satisfied, standard deviation will be used as the substitution variable of exchange rate volatility. In order to overcome the defect of insufficient data because of short time span, average value of exchange rate volatility of 12 months each year will be used. Exchange rate data comes from the IMF. The formula is shown as follows.

$$Vol = \sqrt{\sum_{i=1}^{12} (R_i - \overline{R})^2 / 12}$$

B. Exchange Rate Expectation

Suppose $\theta_t = R_{t+1} / R_t$, then $E(\theta_t) = E(R_{t+1} / R_t) = E(R_{t+1}) / R_t$. Where $E(\theta_t)$ denotes the ratio of expectations to current values of exchange rate. In empirical study, $\widehat{R}_t / R_t$ is used to represent the ratio, where $\widehat{R}_t$ represent the forecast value of the following model: $R_t = a + bt + u_t$, where a is constant and t denotes time trend.

The introduction of $FDIP_{it-1}$ makes (9) becomes the dynamic panel data model. Because of the short time span, the result will be biased if OLS estimation or fixed effect model is used directly. Differential GMM estimation can cause the lack of sample information and reduce the effectiveness of instrumental variables. System-GMM estimation will be used. System GMM estimation can use the information of differential equations and horizontal equations, which benefits to the effectiveness of instrumental variables. Hansen test and AR(·) test will be used to study the effectiveness of instrumental variables and the serial correlation of residual term. Two-step GMM estimation will be used because it is less susceptible to the interruption of heterogeneity. The results are shown in the table below.

TABLE I. RESULTS OF SYSTEM GMM

variables	coefficient	std. error	t-value	Prob.
FDIP(-1)	-0.235	0.006	-39.035	0.000
Vol	-2.329	0.165	-12.872	0.000
R	0.095	0.009	10.896	0.000
E	0.009	0.001	14.157	0.000
Ar(2)_P	0.467			
Hansen_P	0.405			

Note: The table reports the results of System GMM regressions using Stata 11.0. nstrumental variables are two-period lagged R, E, and FDIP.

The result of Hansen test shows that the selection of instrumental variables is effective, residual sequence is uncorrelated by AR(·) test, which means the selection of lagged variable is appropriate. From table 1, we can find that China's outward FDI has a significant delayed effect, both the exchange rate level and the exchange rate expectation of RMB have significantly positive effects on China's outward FDI, but the exchange rate volatility has significantly negative impact on China's outward FDI. Possible reasons are as following. First, the continued appreciation of RMB has made the cost of outward investment lower and lower, for many enterprises, it is better to invest abroad with location advantages than stick to domestic market. Second, the higher the expectation of appreciation is, the lower cost of invest abroad in the future. For higher profits, domestic enterprises will continuously expand the scale of outward FDI. Third, the degree of exchange rate fluctuations means the uncertainty of future cost and revenues, which inhibits the development of outward FDI, because most enterprises are risk averse. Last, the last issue of outward FDI has significantly crowing out effect, because multinational corporations have stable host market share during a certain period.

V. CONCLUSIONS

As the parity rate of various currencies, exchange rate determines the prices of goods and factors of production at home and aboard, affects the scale and flows of outward FDI. In recent years, China's outward FDI has developed rapidly, the appreciation of RMB against dollar and other currencies has provided favorable external conditions to go out. For domestic enterprises, going out will face more vigorous international competition environment, but this is the necessary condition to grow into world-class multinational corporations. Domestic enterprises should go out actively in this era full of opportunities and challenges,

make full use of location advantages and technical advantages of host countries to enhance their international competitiveness.

References

[1] A. Gregory and S. Mccorriston, "Foreign Acquisition by UK Limited Companies: Short and Long-Run Performance," Journal of Empirical Finance, Vol. 1, pp. 99-125, 2005.

[2] V. Sushko, "Foreign Direct Investment under Exchange Rate Uncertainty--Thirty Five Years and Still Uncertain," Advanced International Trade, 2007.

[3] M. Udomkerdmongkol and H. Görg, "Exchange Rates and Outward Foreign Direct Investment: US FDI in Emerging Economies," vol. 4, pp. 754-764, 2009.

[4] D. J. Feils and M. Rahman, "The impact of Regional Integration on Insider and Outsider FDI," Management International Review, vol. 1, 2011, pp. 41–63.

[5] H. Görg and K. Wakelin, "The Impact of Exchange Rate Variability on US Direct Investment," Manchester School, vol. 3, pp. 380-397, 2002.

[6] C. W. Schmidt and U. Broll, "The Effects of Exchange Rate Risk on U.S. Foreign Direct Investment: An Empirical Analysis," Review of World Economics, Vol. 3, pp. 513-530, 2009.

[7] H. Egger and M. Ryan, "Bilateral and Third-Country Exchange Rate Effects on Multinational Activity," University of Bayreuth, 2008.

[8] P. J. Buckley, "The determinants of Chinese Outward Foreign Direct investment," Journal of International Business Studies, Vol. 2, pp. 343-354, 2009.

[9] W. Peiming, D. Joseph and P. Donghyun, "Determinants of Different Modes of FDI: Firm-Level Evidence from Japanese FDI into the US," Open Economies of Review, Vol. 3, pp. 425-446, 2013.

[10] R. Barrell, S. Gottschalk and S. Hall, "Foreign direct investment and exchange rate uncertainty in imperfectly competitive industries," Regionalisation Growth and Economic Integration, 2007.

[11] X.Liangke, "Exchange rate, exchange rate system and FDI, Shanghai Economic Reviw, Vol.10, pp. 25-32, 2012.

[12] S. Gottschalk and S. Hall, "Foreign Direct Investment and Exchange Rate Uncertainty in South-East Asia," international Journal of Finance and Economics, Vol. 4, pp. 349-359, 2008.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Mechanism Innovation and Evaluation Model of Wisdom Tourism under the New Situation Based on Survey of Wuxi

Yawen Cui,Yanping Sun
The Internet of things school
Jiangnan University
Wuxi,214122,China
Email:cywen1208@163.com

Ping Zhu*
College of science
Jiangnan University
Wuxi,214122,China
Email:zhuping@jiangnan.edu.cn

Abstract: Under the "new normal" national construction, wisdom tourism must be chosen in Wuxi to maintain competitive advantages of sustainable development.

By field researches and reviews of the literature, we select three experiences (scenic type, tourism service and tourism expenses) to make the simplified definition of the tourist satisfaction degree. Meanwhile, evaluating factors of this degree including the natural scenery class(A1), cultural relics class(A2), ticket prices(B1), food prices(B2), the price of tourism souvenirs(B3),attractions(C1), tour guide service(C2),catering(C3),transportation(C4), entertainment(C5) and shopping(C6) are abstracted. In order to verify the assessment system by evaluating tourist satisfaction degrees of five scenic spots (Ling mountain, Turtle Head Islet, the Three Kingdoms and Water Margin, Li Park and former residence of Fucheng Xue) in Wuxi in macro-level and further analyze evaluating factors in micro-level, we establish the Assessment System of Tourist Satisfaction based on Analytic Hierarchy Process model and the Importance-Satisfaction of Evaluating Factors model. According to the analysis on results of the evaluation above, reasonable and effective suggestions are put forward which direct the tourism development in Wuxi and it can also make a good theoretical foundation for the innovation of mechanism design for the wisdom tourism and provide effective and practical data.

Key Word: wisdom tourism in Wuxi, Tourist satisfaction model, Importance-satisfaction of factors model, the new normal.

I. THEORETICAL SOURCE AND CONSTRUCTION OF THE MODEL

A. Connotation of tourist satisfaction

Wang Xia [1] thinks that tourist satisfaction is the result of comparison between tourist expectation and field travel perception and it stresses the comparison process of tourist and its results. Wang Kai [2] thinks that the index of tourist satisfaction has four stages: expectation satisfaction, experience satisfaction, evaluation satisfaction and post-tourism satisfaction. After that, he builds a tourist satisfaction evaluation model and makes the empirical analysis and analyzes it with examples.

B. Evaluation factors system of tourist satisfaction

According to a large number of literatures, we create a tourist satisfaction model based on analytic hierarchy process to evaluate the tourism in Wuxi scientifically. We divide the evaluation factor system of tourist satisfaction into three layers: the objective layer, the project

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.39

layer and the factor layer. We utilize analytic hierarchy process (AHP) to determine the weight of evaluation indexes. The contents of each layer are shown as follows.

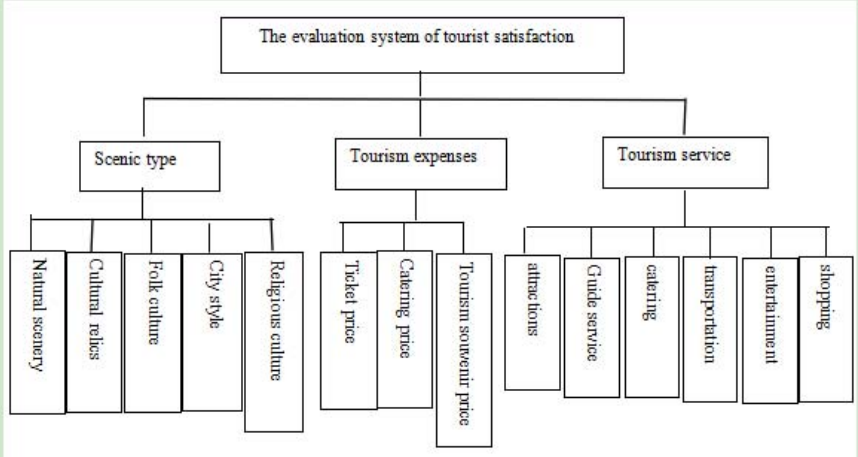


Figure 1. Evaluation factors system of tourist satisfaction

II. MODEL

A. Tourist satisfaction model based on analytic hierarchy process

$$SAT_p = \sum_{f=1}^{n} w_f * x_f \qquad [3]$$

$$SAT = \sum_{j=1}^{n} w_p * SAT_p .$$

Where w_f is defined as the weight of factor *f*. x_f is the mean satisfaction of factor *f*. SAT_p is the mean satisfaction of project *p*. w_p is the weight of project *p*. x_p is derived from results of questionnaire analysis. w_f, w_p are derived from programming by Matlab.

TABLE I. EVALUATION FACTORS SYSTEM OF TOURIST SATISFACTION

objective	project	weight	factor	weight
Tourist satisfaction(A)	Scenic type(B1)	0.3753	Natural scenery(C11)	0.4689
			Cultural relics(C12)	0.2605
			City style(C13)	0.1563
			Folk culture(C14)	0.0621
			Religious culture(C15)	0.0521
	Tourism expenses(B2)	0.3326	Ticket price(C21)	0.4286
			Catering price(C22)	0.3333
			Tourism souvenir price(C23)	0.2381
	Tourism service(B3)	0.2920	Attractions(C31)	0.3333
			Guide service(C32)	0.0370
			Catering(C33)	0.2963
			Transportation(C34)	0.1852
			Entertainment(C35)	0.1111
			Shopping(C36)	0.0037

B. Factor Importance-Satisfaction Model [3]

In order to reflect more directly, the results of the tourist satisfaction measurement are expressed in importance - satisfaction matrix distribution model. Horizontal axis represents the importance of evaluation factors while vertical axis represents the satisfaction of evaluation factors. 4 quadrant regions in Matrix model represent the district of advantages promoting (high importance and high satisfaction), the district of maintaining (low importance, high satisfaction), the district of subsequent opportunity (low importance, low satisfaction), and the district of urgent improvement (high importance,) low satisfaction in the development and the construction of the geological park.

III. DATA PROCESSING AND RESULTS ANALYSIS

A. The processing of important data.

According to the requirements of our project, we collect about 3000 questionnaires in different times (work day, weekend, holiday), places (different types of scenic area: toll scenic spot and no charge scenic spot) and research objects. After that, about 1949 valid questionnaires are extracted. Given that five scenic spots that we selected are scenic spots with charge and our objective is derived from satisfaction data oriented to the outsiders, we filter the research data collected from local people.

B. Results and Analysis of models

In general, the scores of the 5 scenic spots are between 1.5 and 1.9, and the difference is small. The factors concentrate on the district of maintaining and the district of advantages promoting. Natural scenery, cultural relics, ticket price and catering price lie in the district of urgent improvement. Attractions and catering service lie in the advantage promoting district. Transportation, entertainment, guide service and shopping lie in the maintenance district and tourist souvenir price lies in the district of subsequent opportunity. Specific analysis is shown as follows.

Model one: Tourist service Project gets the highest score. The scores in a descending order are Ling mountain, the former residence of Xue Fucheng, Three Kingdoms Water Margin, Li park, Turtle Head Islet. Tourism expenses Project are the second one. The scores in a descending order are the former residence of Xue Fucheng, Li park, Turtle Head Islet, Three Kingdoms Water Margin and Ling mountain. The scores of Scenic type Project in five scenic spots differs largely. The scores of Ling mountain and Turtle Head Islet are far more than that of Three Kingdoms Water Margin, Li park and the former residence of Xue Fucheng.

Model two: Taking Ling mountain which receives the highest satisfaction degree as the base. The satisfaction of ticket price is higher than that of Natural Scenery and cultural relics in Li park. The satisfactions of cultural relics are higher than natural scenery in Three Kingdoms Water Margin while they are all lower than those in Ling mountain. The satisfactions of Ticket Price, catering price and tourist souvenir are much higher than that of natural scenery and cultural relics. The satisfaction of natural scenery in Turtle Head Islet is higher than that in Ling mountain while ticket price in Turtle Head Islet is lower than that in Ling mountain.

IV.CONCLUSIONS

Scenic type is the key factor of the tourist satisfaction, but the corresponding score in five scenic spots are all not in the highest level. Natural scenery class and the class of cultural relics are all in the district of urgent improvement. The scenic spots can develop some special tourist products, such as DIY

products or leisure tourism products.

Tourism service is the core factor to affect tourist satisfaction. The satisfaction score of five scenic spots are all in the highest level. Two factors (catering and attractions) lie in the district of advantages promoting while transportation, entertainment, tour guide service and shopping lie in the district of maintaining. Our advices are represented as follows: establish efficient employee evaluation system to advance employee management; strengthen the service management in dining and shopping to improve qualities; increase tourism lines with narrators; launch self-driving tour products.

Tourism spending is an essential condition to support all kinds of tourists' activities. Ticket price and catering price lie in the district of urgent improvement while the price of the tourism souvenir lies in the district of subsequent opportunity. It is essential to lower the ticket price advisably, enhance the satisfaction degree by shopping and entertainment by referring to the tourism industries in Taiwan.

REFERENCES

[1]Xia Wang,Zehua Liu,Hong Zhang, Review and Prospect of the study on tourist satisfaction.Journal of Beijing International Studies University. Vol 1.pp:22-29.2010

[2]Kai Wang,Chengcai Tang,Jiaming Liu, Tourist satisfaction index evaluation model of Cultural creative tourist attractions taking Beijing 798 Art District as an example.Journal of travel.vol 26.pp:36-44.2011

[3]Feng Wang, Comprehensive evaluation and sustainable development study on Taihu ecological agriculture tourism circle.Nanjing: Nanjing agricultural university.2010

Fund Project:Project supported by the National Natural Science Foundation of China(61300150,71271029) National college students innovative training program(201410295045)

* Communication author:Ping Zhu,professor of College of Science.

Design and Implementation of Logistics Information Management System Based On Web Service

Hua Jiang
Department of Information Shandong Science Technology Vocational College
Weifang ,Shandong province 261053,China
Email: wf-jianghua@163.com

Yuman Li
School of Electrical Engineering, Chongqing University
Chongqing, Sichuan province 400030,China
Email: mandylym@foxmail.com

Hua Fang
Department of Management & Information Shandong Transport Vocational College
Weifang ,Shandong province 261206,China
Email: sdfanghua@sina.com

***Abstract*—the paper provides the software architecture, design method and implementation techniques for a Web service-based logistics information system which is based on Browser, Web Server, Application Server, Data Structure. An example is given to illustrate how to fulfill the communication between the server and the client.**

Keywords-component; Web Service; Logistics; Management System

I. INTRODUCTION

With the development of computer and network technology, the modern logistics information technology has become a new hot spot of China's economic development. With the deepening of logistics information, the software and the internal systems of the modern logistics enterprises have become more complex, such as online inquiry system, electronic office systems and financial management. These systems are cross-cutting. A lot of information and data are reused. But these systems can not be smooth information exchange and sharing with each other.

Through systematic analysis and the actual design of the logistics management system, a feasible ways and means is provided for information technology companies. During development of the logistics management software system, using the Web Services, ASP.NET, and other functions which are provided by Microsoft platform, some technical problems are solved.

II. THE KEY TECHNOLOGY

A. *ASP.NET*

ASP.NET is a new powerful technology, which is used to write dynamic Web pages. ASP.NET is a combination of Microsoft's ASP and the .NET technology. The aims of Microsoft are to reform the methods of program development and the methods of the companies engaged in business activities in the future. Therefore, as the combination of ASP and .NET, ASP.NET is a way to create dynamic Web pages.

ASP.NET is built on .NET Framework classes, and provides "Web Application Templates" composed of the controls and basic part. It greatly simplifies the development of Web applications and XML Web services. Programmers are faced with a set of direct ASP.NET controls, such as textbox, button, which are packaged by common HTML user interface component. In fact, these controls run on the Web server, and send the user interface to the browser in the form of HTML.

B. *Web Services*

Web Services is a dynamic and interoperable environment deployed on the Web, which is built by objects, components, network access, and combinations of multiple applications. It is a technology architecture, which is designed to solve call and integration between the loosely coupled client and Web Services, between Web Services in a network environment.

Web Services is the generic term for a range of technologies, including XML, SOAP, WSDL (Web Services Description Language, called WSDL), UDDI (Universal Discovery, Description, and In-victory, referred UDDI). Generally, Web Services is an application as a service offering. And it is a resource which is targeted by URL, automatically return information to its clients. As the foundation of Web Services technology, XML is the standard technology to descript data and information in an open environment. SOAP is a lightweight protocol to exchange information in a distributed environment. It inherits the openness and description scalability of XML，and support SMTP, FTP, TCP and other standard network protocols. UDDI is a standard based on the web to distribute register, publish and discover web Services. It describes the call interface provided by web Services. WSDL is an XML-based component description standard. It is described Web Services and its functions, parameters, return values and other information.

The architecture of Web Services is based on the interaction of three roles (service providers, service registry and service requestor). Interaction is completed through publish, find and bind. Together these roles and operations act on the Web Services components. Therefore, the

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.40

interaction of these three roles involve publish, find and bind operations:

- The service provider primarily publishes Web services.
- The registry service is equivalent to Query Center, all published Web services can be find the appropriate records at here.
- The service requestor obtain the location of the owner of Web services and related information by querying the service registry center, and complete the desired action through standard call interface to interact with a service provider etc.

III. SYSTEM DESIGN AND IMPLEMENTATION

A. System Development Environment

In the logistics management system development environment, the server operating system hard disk parts to NTFS format. And the following development tools are installed during the software development process:

- During the process of database development using SQL Server2008 database is used.
- In the Web application development process, the Microsoft Visual Studio NET 2008 integrated development environment is used as software development tools.
- In the Web application development process, the .NET Frame-work SDK development kit is used as ASP.NET development environment.
- In the Web application development process, the Microsoft Internet Information Services (IIS) 6.0 is used as a Web server, and Microsoft Internet Explorer is used as a Web browser.

B. System Architecture

System is divided into three layers: the user interface layer, business layer used SOAP to access, database layer. We use the three layer structure is mainly to make the project structure more clear, more clear division of labor, the latter's maintenance and upgrading more favorable.

The following diagram shows the relationship between the three layer structures.

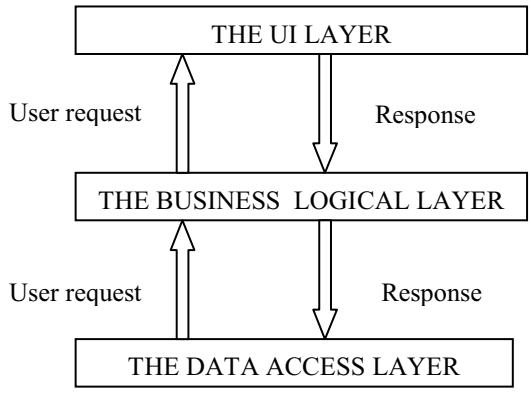


Figure 1. The three layer structures

The user interface layer (UI) is composed by the JSP page. It communicates with the business layer through the interface provided by the system. It submits the various data entered by the user and form data to the business layer, and accepts the results or the response returned from business layer. At last, the user interface layer provides various services in the form of Web pages to users.

The function of the business logical layer (BLL) is communication between the presentation layer and the data layer. It processes the data received from the presentation layer or database operations, and submits query or modifies query to the database. And it putts the results of the appropriate action to the presentation layer. The business layer uses SOAP as a protocol to develop. It shields the details of the database operation.

Database access layer (DAL) is responsible for data management. Its function is mainly responsible for the database access. The simple statement is to achieve the data table Select, Insert, Update, Delete operation. Through this layer, the data which stored in the database will be submitted to the business layer, and the data which accessed by the business layer shill be stored in the database.

All of these are based on the UI layer. The user's needs are reflected to the interface (UI), UI is reflected to the BLL, the BLL is reflected to the DAL, the DAL is operated by the data, and the operation is returned to the user until the user needs the data to be returned to the user.

C. Function module

The function module of logistics information system mainly includes the following main subsystems: business management, warehouse management, distribution management, accounting management and customer service. Each subsystem is described as follows:

- Business management subsystem is the fundamental of the logistics system. It provides the business to customers, orders and other basic information processing.
- Warehouse management subsystem can achieve centralized management of warehouse resources including the different regions, different attributes, different specifications, and different cost.
- Distribution management system in accordance with the principle of just in time distribution, to achieve production enterprises zero inventory production raw material distribution management, and commercial enterprises small batch and multi varieties and distribution management, realize the common distribution and multi echelon distribution management.
- The billing model provided by the accounting management subsystem is highly structured. First logistics center will all billing project to establish data base, by the packing of the goods customer service personnel according to the customer, goods attributes, customer service requirements of valuation price, until the customer signed a contract after the entry billing system, after each time the customer's goods were all operations, billing system

will automatically according to the charging unit billing, without any manual input. When the arrival of the settlement date of the customer, the settlement will be automatically carried out, and then to the financial system in accordance with the accounting subjects, for financial personnel and the customer's account.

D. System development

The development of Web Service consists of server-side program development, deployment services, client development etc. The implementation of the work order of the system is taken as a sample.

A) Workflow of System

- Users can access to the system from home though entering their user name and password.
- The LoginService service is called to verify the legitimacy of user name and password.
- The system page is entered if the validation is successful. And then GetUserItem method of GetItemService is called to query the user's to-do items in the system.
- The user can choose to deal with to-do project, other services or exit the system.
- If the user chooses to deal with project, the GetItemByID method of the work orders services is called to obtain detailed information. The project is processed and the result is returned.
- The internal process of the system work is controlled by the workflow system of logistics companies.

B) Realization of GetItem Service

```
/* define GetItemService interface class */
public interface GetItemInf{
public Item[] GetUserItem (String username)
throws ItemException;
public String GetItemByID(int id, int Sid)
throws ItemException;
……
}
/*define GetItemService implementation class*/
public class GetItemService implements GetItemInf{
public GetItemService() throws ItemException{}
public Item[] GetUserItem (String username)
throws ItemException{……}
public String GetItemByID(int id, int Sid)
throws ItemException{……}
……
}
```

C) Web Service Deployment

After Web Service server-side program development is completed, it needs to be deployed on the WEB server for client calls. There are two methods can be used to deploy a Web Service to be registered on the server.

The first method is to register in the form of GUI with the direct use JSP pages provided by ApacheSOAP. Start the tomcat, go to the path to the soap server configuration, open your browser and enter the URL, it will display the soap home page. Then click Run hyperlinks, the soap service management page is opened. At last, click the deploy button, input ID, Methods, such as content in the page, and then click the deploy button at the page button. The deployment of the service is finished.

The second method is to register the appropriate service through the deployment descriptor. This method uses tool class which is provided by ApacheSoap framework: org.apache.soap.server.ServiceManagerClient to deployment descriptor to bind to a soap request, and the request is submitted to the server.

D) Client implementation

After deploying Web services on the server, Web service deployed on the client can be called. Take invoking GetUserItem method of GetItemService services as an example to illustrate the client is how to call a server-side SOAP services.

```
Callcall= new Call();
/* create a Call object, use the object to complete the call to the SOAP */
Parameterparam=new Parameter ( );
/* parameters passed to the Call object */
call.setTargetObjectURL(uri);
call.setM ethodName(remoteM ethod);
call.setParameter( );
SOAPM appingRegistry smr=new SOAPM appingRegistry();
……smr.mapTypes( );
/*using the mapTypes method provided by the appingRegistry SOAPM object to create a mapping between the Java object and the XML */
call.setsSOAPM appingRegistry(smr);
/* the map is assigned to the Call after it is established */
Responseresp = call.invoke();
/* Call the object's invoke () method to invoke the SOAP service */
```

E. System security control

Before the implementation of the logistics information system, the security control of system network, hardware, software and data is very important. This system mainly carries on the control design to the information system of the web, the system access, the database and so on. The following example of system access is described.

Registered users are authorized to access the system's users. The user role is assigned for each user, through by the different of the user's business. Each role has different functions and database access rights. For example, a number of roles allow to view data but not to modify the data. So, some functions and interfaces of the system are shielded by different types of roles, that is, the system is configured on the user interface and data.

IV. CONCLUSION

Due to the logistics information management based on Web Services with a good cross-platform, extensive integration flexibility, fast and efficient and scalable, so logistics management system based on Web Services must

be the new trend of development traditional C / S mode. The Web Services will play an increasingly important role in the system implementation. The project achieved the combination of logistics management system and Web Services technology, discussed the design and development of personalized enterprise management systems, and provided a new way of thinking of promoting the information of construction.

REFERENCES

[1] SVinoski.Integration with Web Services[J].IEEE Internet Computing,2003,7(6):75-77.

[2] Jun Zhu. Web Services Provide the Power to Integrate[J].IEEE Power and Energy Magazine,2003,1(6):40-49.

[3] Dongsao Hao, Jongyong Goo .Design of a Web Services Based eAI Framework[J].IEEE Advanced Communication Technology,2004,6(2):1003-1008.

[4] M C Carboneras,C M Insa,E V Salort. ERP Implementation in the Stone Industry Special Difficulties and Solutions in the Production Area[J].IEEE Emerging Technologies and Factory Automation,2003,2(2):146-149.

Study of Copyright Protection for Merchandise Pictures in E-Commerce

Liyi Zhang
School of information management
Wuhan University
Wuhan，China

Chang Liu
School of information management
Wuhan University
Wuhan，China
liuchang0310@163.com

***Abstract*—For solving the picture misappropriation problem in e-commerce, this paper proposes a content-based copyright detection method for e-commerce website pictures. First, a picture is smoothed by bilateral filter in order to effectively preserve the key points in the edges and low-contrast regions, which ensure the integrity of its key points. Second, the picture's texture feature is extracted by using Gabor filter, and the shape feature is captured by Hu moment. Third, feature vectors are assigned weights and normalized in order to reduce the effects of variations in their dimensions and components. Finally copyright-infringing pictures are detected by calculating the Euclidean distance and we use fuzzy heuristics to measure similarity between the query and the database images. The test dataset is crawled from the largest e-commerce website (Taobao.com). The copyright-infringing pictures can be detected by the method and the average accuracy reaches 91%, achieved the desired effect.**

Keywords- e-commerce; picture misappropriation; copyright detection; fuzzy inference; picture features

I. INTRODUCTION

Presently, China's e-commerce is in a state of rapid growth. By the end of 2014, its gross merchandise volume (GMV) had reached US$470 billion, and China has surpassed the U.S. as the world's largest online retail market. However, counterfeit products and false advertising hamper the further growth of e-commerce. Some merchants have misappropriated and abused merchandise pictures belonging to other merchants or websites in advertising their own counterfeit products. For solving the problem, this paper proposes a content-based copyright detection method for pictures.

Currently, digital watermarks are the most commonly method used for copyright protection of pictures. However, many deficiencies of digital watermarks have been noted. First, watermarks must be embedded prior to publication, and copyright protection or detection is impossible for unmarked works. Second, watermarks are vulnerable to hacking. Once cracked, a watermark system no longer offers any protection. For these reasons, content-based image protection has also been studied. The visual features of an image, such as shape, color, and texture, are directly related to its content[1]. Lowe[2]utilized point feature for duplicate picture detection and sub-graph retrieval. Lin C H et al. [3]proposed a smart CBIR application based on color and texture features, but it is not effective against deformed (e.g. rotated) images. ElAlami[4] extract the implicit knowledge from the image data by carrying-out clustering and set the rules based on the relevance feedback given by experts in order to refine the results by improving clusters. Kim[5]proposed a copy detection method based on block DCT coefficients, but this method of feature extraction is dependent on image blocking, and weak against common geometric attacks such as cutting, scaling and translation. Syam B and Rao Y S[6] proposed a GA-based similarity measure in CBIR. The growth of e-commerce has given new applications for theories and technologies of computer visuals, including that of picture retrieval. Chen H et al.[7]developed an automated system that uses semantic properties to describe the features of garments, extract visual features of garment types, and formulate garment style rules based on relations between these properties. Liu S et al.[8] provided a solution to garment retrieval across scene, allowing for retrieval of similar products in online stores based on street pictures. Bossard et al.[9]created a system that allows for classification and description of upper body garments. Lo C C et al.[10] proposed a recommendation system named the Mobile Merchandise Evaluation Service Platform (MMESP), which allows using product pictures for real-time product identification.

II. PROBLEMS IN IMAGES PROTECTION

For protecting the interests of merchants, digital watermarks are used in Taobao to image copyright protection. Merchants should upload the images to apply for certification, and use the certificated images to show their products. But it still exists some limitations. First, the requirement to the image is quite high. There can be no stitching and text in the picture, and the picture should not be too simple and have no visual decoration. Currently, it can only be used in dress pictures. Second, it is vulnerable to hacking. Once cracked, the system no longer offers any protection.

In their misappropriation of pictures, merchants would often manipulate the original pictures in various ways, such as translation, rotation, cutting and pasting, insertion of text, removal of picture elements, and color changes. SIFT (Scale Invariant Feature Transform) is capable of matching pictures that have undergone translation, rotation and other deformations. Shape features is invariant to location, orientation and translation, and can reflect the characteristics of image effectively. Texture refers to innate surface properties of an object and their relationship to the

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.41

surrounding environment. It is measured by the relative brightness but will still be affected by the image quality.

We present a copyright detection approach, which is based on texture, shape and SIFT features.

III. METHODOLOGY

A. SIFT Feature Extraction Based on Bilateral Filter

Fundamentally, SIFT involves finding extreme points in the scale space, and extracting their invariants for position, scale and rotation. Traditional SIFT applications have used Gaussian filter for picture smoothing. Their main weakness is their tendency to remove key points at the edges and in low-contrast regions after detection. This is due to their selection mechanisms for key points. In general, SIFT tends to select for points in the corners and high-contrast regions, and disregard key points from the edges.

Bilateral filter is a non-linear filter that achieves edge preservation and noise reduction by taking account of both spatial proximity and pixel value similarity, while accommodating for spatial information and gray-scale similarity simultaneously[11][12]. It is an adaptive Gaussian filter notable for being simple, non-iterative and localized.

We used bilateral filter to replace the Gaussian filter for picture smoothing and generation of scale space. A bilateral filter uses pixel locations and brightness gradients to control the weight values. Points around regions of great brightness variations are given low weights for edge preservation, while regions of small brightness variations are given normal Gaussian weight values for smoothing.

Let $f(x,y)$ be the input picture, and $g(x,y)$ be the output picture after smoothing. The smoothing process of bilateral filter can be expressed as:

$$g(x,y)=\frac{\sum_{n=-w}^{w}\sum_{m=-w}^{w} f(x+m,y+n)\exp(-\frac{m^2+n^2}{2\sigma_1^2})\exp(-\frac{(f(x,y)-f(x+m,y+n))^2}{2\sigma_2^2})}{\sum_{n=-w}^{w}\sum_{m=-w}^{w}\exp(-\frac{m^2+n^2}{2\sigma_1^2})\exp(-\frac{(f(x,y)-f(x+m,y+n))^2}{2\sigma_2^2})} \quad (1)$$

Where σ_1 is the parameter that control Gaussian shape in space; σ_2 is the parameter to control the effects of brightness changes. When σ_1 increases, the filter will cause a stronger blurring effect. When σ_2 decreases, the edges will be preserved. The purpose of using bilateral filter is to protect the edges, and ensure more key points along edges can be detected.

For each picture, its gradient magnitude $M(x,y)$ and orientation $O(x,y)$ can be expressed as:

$$M(x,y)=\sqrt{(\frac{\partial g(x,y)}{\partial x})^2+(\frac{\partial g(x,y)}{\partial y})^2}$$

$$O(x,y)=\arctan\left(\frac{\frac{\partial g(x,y)}{\partial y}}{\frac{\partial g(x,y)}{\partial x}}\right) \quad (2)$$

Figure 1 shows the picture scale spaces generated after smoothing by bilateral and Gaussian filters, with the left using bilateral filter and the right using Gaussian filter.

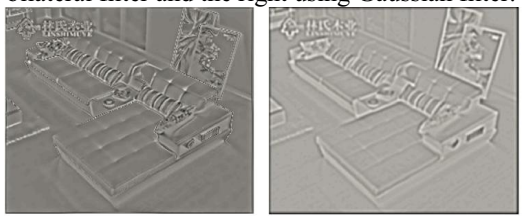


Figure 1. Comparison of picture scale spaces generated by Bilateral filter and Gaussian filter

The edges are better preserved in the left. Generally speaking, values output by Difference of Bilateral are at a lower order of magnitude compared to Difference of Gaussian. Several obvious extremes can be seen in the picture processed with Difference of Bilateral, which are not found in the picture processed with Difference of Gaussian, as these values have been filtered out by the Gaussian filter.

Figure 2 shows the number of key points generated after smoothing by bilateral and Gaussian filters.

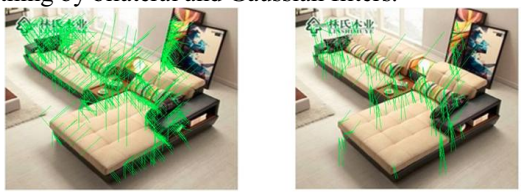


Figure 2. Comparison of number of key points generated by Bilateral filter and Gaussian filter

More key points are visible in the left, generated by the bilateral filter, than the right, generated by the Gaussian filter.

Figure 3 shows a comparison of key point matching with pictures smoothed by Gaussian and bilateral filters.

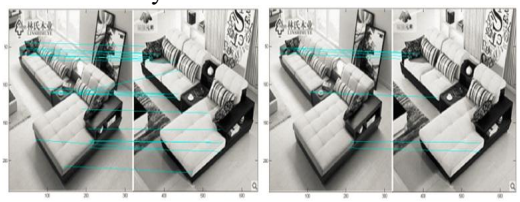


Figure 3. Comparison of matching results using Bilateral filter and Gaussian filter

There are notably more matching key points from using the bilateral filter than the Gaussian filter. This is due to the edge protection effect of bilateral filters, which increased the amount of extracted information, and thus increased the number of matching features.

B. Extraction of Shape Feature

Hu moment which is invariant to location, orientation and translation is widely used in the area of image

classification. The moment of order $(a+b)$ of an picture $I(x,y)$ is defined as: $m_{a,b}=\sum_x\sum_y x^a y^b I(x,y)$, where $a,b=0,1,2...$,the sum are over the values of the spatial coordinates x and y spanning the picture. The corresponding central moment is defined as: $u_{a,b}=\sum_x\sum_y (x-\bar{x})^a (y-\bar{y})^b I(x,y)$, where $\bar{x}=\frac{m_{10}}{m_{00}}$, $\bar{y}=\frac{m_{01}}{m_{00}}$, which are referred to as region center. The scale invariant moments $\eta_{a,b}$ where $(a+b)\geq 2$ can be defined as: $\eta_{a,b}=\frac{\mu_{a,b}}{\mu_{0,0}^{\gamma}}$, where $\gamma=[\frac{a+b}{2}]+1$. Then seven shape features ϕ_1 — ϕ_7 which are scaling, rotation and translation invariants can be extracted.[13]

C. Extraction of Texture Features

The texture feature of images is commonly represented by co-occurrence matrix, which is based on directions and distances between pixels. Gabor filter is the most effective and widely used texture analysis method.

Gabor wavelet function is sensitive to image edges, provides excellent direction and scale selectiveness, and is also insensitive to illumination changes, and tolerant to a certain degree of image rotation or deformation. Hence we utilized Gabor filter to extract texture features.

When a picture is processed by Gabor filter, the input is a convolution of picture $I(x,y)$ and Gabor function $g(x,y)$. After applying Gabor filter on scale and orientation, an array can be obtained[14]:

$$E(m,n)=\sum_x\sum_y |G_{mn}(x,y)| \qquad (3)$$

$$m=0,1,...,M-1; n=0,1,...,N-1$$

The magnitudes represent the energy content at different scale and orientation of image. Texture-based image retrieval allows find pictures or regions with similar textures. The texture feature of regions are expressed using mean $\mu_{m,n}$ and standard deviation $\sigma_{m,n}$:

$$\mu_{m,n}=E(m,n)\Big/P\times Q$$
$$\sigma_{m,n}=\sqrt{\sum_x\sum_y (|G_{mn}(x,y)|-\mu_{m,n})^2\Big/(P\times Q)} \qquad (4)$$

Where M represents the scale; N represents the orientation; P and Q represent the height and width of the input image. The resulting mean $\mu_{m,n}$ and standard deviation $\sigma_{m,n}$ constitute two feature vectors respectively, which are then combined into a single feature vector as the texture descriptor.

D. Similarity Detection

Before the feature vectors are combined, they are normalized to reduce the effects of different feature dimensions and variances of the feature components. The normalized feature can be represented as:

$$F_D=\left[\omega_s\times\frac{f_s}{N_s\cdot\delta_s\mu_s},\omega_t\times\frac{f_t}{N_t\cdot\delta_t\mu_t}\right] \qquad (5)$$

Where N_s and N_t are the dimensions of shape and texture feature vectors; δ_s , δ_t and μ_s , μ_t are the average values and standard deviations of shape and texture; ω_s and ω_t are weights of shape and texture; $0\leq\omega_s,\omega_t\leq 1$ and $\omega_s+\omega_t=1$.

Once the feature vector is extracted, a similarity detection method can be used to compare the input image against images in the database. The choice of similarity formula is vital to the process. We chose Euclidean distance for our similarity detection and calculate visual feature distance as similar standards:

$$d(F^Q,F^T)=\sqrt{\sum_{i=0}^{n-1}(F_i^Q-F_i^T)^2} \qquad (6)$$

Where n is the dimensions of the feature vectors, F^Q , F^T are feature vectors of the input picture and an picture from database.

For detecting the illegal images or the images similar to the original images, fuzzy inference method is used to detect the similarity. Shape features and texture features are two visual features of images that have been retrieved using Moment Invariants and Gabor wavelet. The first priority is given to the shape features, as shape of an image is not easily affected by external factors, and also it is invariant to the rotation, translation and orientation. The second priority is given to the texture features. The performance will be improved by defining these criteria along with the fuzzy rules. The Mamdani fuzzy inference method is used to perform fuzzy rules in our proposed approach[15].

We introduce a set of fuzzy rules to process the results achieved by applying the two algorithms discussed above. (1) We define a number of inputs. There are two inputs which are shape distance, and texture distance between query image and database images in our approach. (2) The membership functions for two types of input have been defined. The types of fuzzy set that identified each input as low, medium and high. Three types of output fuzzy sets have been declared such as high similar, medium similar and low similar. (3) A fuzzy rule can be defined as a conditional statement. Fuzzy rules applied using logical operator. To process the Mamdani fuzzy inference method. We take the crisp inputs and fuzzy them to determine the degree to which these inputs belong to each of the appropriate fuzzy set. The AND fuzzy operator is applied to get one number that represents the result of antecedent of rules. (4) The above fuzzy rules are used for data aggregation. (5)The aggregate output fuzzy set should transform to a single crisp number.

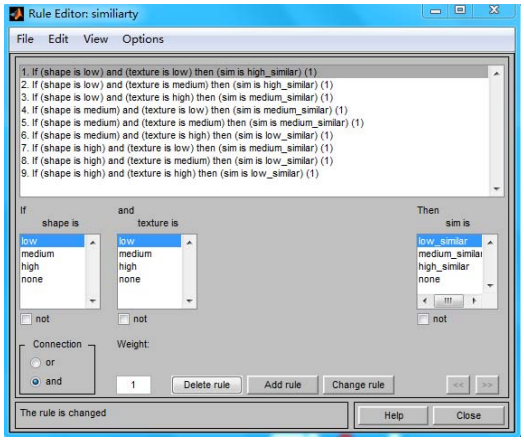


Figure 4. A set of fuzzy rules applied to priorities results

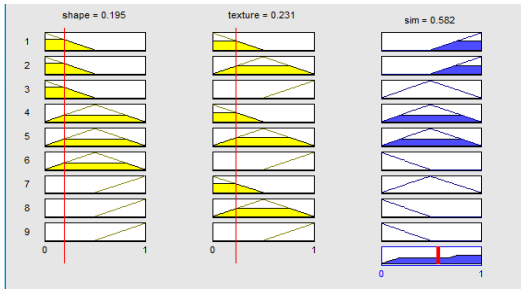


Figure 5. Fuzzy rules representation

IV. RESULTS

For verifying the feasibility of the method, we created a database using pictures crawled from Taobao. We searched for the key word on Taobao and stored the 1000 pictures from the first 20 pages. As the method is based on content-based image retrieval systems, it would be compared to the methods proposed by ElAlami[4] and Wang X Y et al.[14].

In the 3 figures, the first picture is the query picture which the database pictures are compared against. Figure 6 shows the content-based detection method is capable of identifying pictures that are modified based on the query picture. The results of Figure 6 are a notable improvement on Figures 7 and 8, which are closer to each other. In e-commerce sites, for differ from the original pictures, people will make some changes to color using some technical measures. Therefore, the accuracy is not good enough when detect the copyright using color features. The measures proposed by ElAlami and Wang X Y et al. used color and texture features to retrieve pictures, for the above-mentioned reasons, the results are not good enough. Considering the practical situations in e-commerce site, we use texture, shape and SIFT features to detect the similarity, and the result is better than the results using the measures proposed by ElAlami and Wang X Y et al. For a more direct comparison, the results can be described using precision and recall.

Figure 6. Results of our method

Figure 7. Results of method proposed by ElAlami

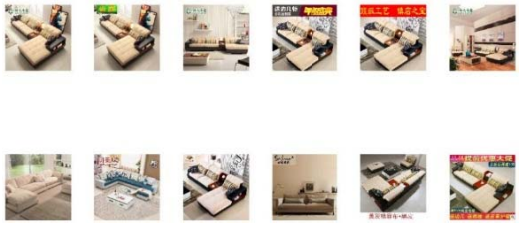


Figure 8. Results of method proposed by Wang X Y et al.

The most commonly used indicators for evaluation of image retrieval systems are precision and recall. Precision represents the ratio of retrieved pictures relevant to the query, and recall represents the ratio of relevant pictures that have been retrieved. They can be defined as[14]:

$$precision = \frac{a}{a+b}$$
$$recall = \frac{a}{a+c} \quad (7)$$

Where a is the number of relevant pictures retrieved, b is the number of irrelevant pictures retrieved, and c is the number of relevant pictures that have not been retrieved. The precision and recall ratios are then combined into a single parameter F-score:

$$Fscore = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

The performances of three retrieval methods are calculated for comparison below.

TABLE 1. COMPARISON OF OUR METHOD VERSUS TWO EXISTING METHODS

	Precision	Recall	F-score
Our method	0.91	0.63	0.74
Method proposed by ElAlami	0.72	0.33	0.45
Method proposed by Wang X Y et al.	0.75	0.39	0.51

It is obvious that the method introduced in this study performed significantly better than the two existing methods. The other two methods had similar performances, with the latter slightly better than the former.

V. Conclusion

In view of the picture misappropriation problem in e-commerce, this paper proposes a content-based copyright detection method. Copyright-infringing pictures can be detected by the method. We conducted tests using pictures crawled from Taobao, China's largest e-commerce website, which resulted in an average accuracy of 91%. A comparison of results using precision and recall as indicators for evaluation shows that the method had significantly better performance than existing methods, and greatly improved the results of detection. More work should be done in order to improve the efficiency of detection and make the method more practical for databases with massive amounts of pictures. In addition, it is a common practice in the industry for multiple licensees to share the use of pictures provided by one distributor. Removing this type of pictures from detection results will further improve the accuracy.

References

[1] K. Shkurkol and X. Qi, "A RADIAL BASIS FUNCTION AND SEMANTIC LEARNING SPACE BASED COMPOSITE LEARNING APPROACH TO IMAGE RETRIEVAL," in Acoustics, Speech and Signal Processing, 2007, pp. 945–948.

[2] S. Keypoints and D. G. Lowe, "Distinctive Image Features from," vol. 60, no. 2, pp. 91–110, 2004.

[3] C.-H. Lin, R.-T. Chen, and Y.-K. Chan, "A smart content-based image retrieval system based on color and texture feature," Image and Vision Computing, vol. 27, no. 6, pp. 658–665, 2009.

[4] M. E. Elalami, "A novel image retrieval model based on the most relevant features," Knowledge-Based Systems, vol. 24, no. 1, pp. 23–32, 2011.

[5] C. Kim, "Content-based image copy detection," Signal Processing: Image Communication, vol. 18, no. 3, pp. 169–184, 2003.

[6] B. Syam, Y. S. Rao, S. Associate, and A. Pradesh, "An effective similarity measure via genetic algorithm for Content Based Image Retrieval with extensive features," INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY, vol. 10, pp. 143–151, 2013.

[7] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7574 LNCS, no. PART 3, pp. 609–623, 2012.

[8] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3330–3337, 2012.

[9] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7727 LNCS, no. PART 4, pp. 321–335, 2013.

[10] C.-C. Lo, T.-H. Kuo, H.-Y. Kung, H.-T. Kao, C.-H. Chen, C.-I. Wu, and D.-Y. Cheng, "Mobile merchandise evaluation service using novel information retrieval and image recognition technology," Computer Communications, vol. 34, no. 2, pp. 120–128, 2011.

[11] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," Sixth International Conference on Computer Vision (IEEE Cat No98CH36271), 1998.

[12] M. Tanaka and M. Okutomi, "Latent common origin of bilateral filter and non-local means filter," Image (Rochester, NY), vol. 7532, pp. 753202–753202–10, 2010.

[13] MING-KUEI HU, "Visual Pattern Recognition by Moment Invariants"," IRE Transactions on Information Theory, vol. 8, no. 2, pp. 66–70, 1962.

[14] X.-Y. Wang, H.-Y. Yang, and D.-M. Li, "A new content-based image retrieval technique using color and texture information," Computers & Electrical Engineering, vol. 39, no. 3, pp. 746–761, 2013.

[15] D. R. Keshwani, D. D. Jones, G. E. Meyer, and R. M. Brand, "Rule-based Mamdani-type fuzzy modeling of skin permeability," Applied Soft Computing Journal, vol. 8, no. 1, pp. 285–294, 2008.

Applying Innovation Resistance Theory to Understand Consumer Resistance of Using Online Travel in Thailand

KANJANA JANSUKPUM
School of Information Management
Wuhan University
Wuhan, China
E-mail: kjn986@yahoo.com

SUPAMAS KETTEM
Faculty of Humanities and Social Sciences
Phetchaburi Rajabhat University
Phetchaburi, Thailand
E-mail: skettem@gmail.com

Abstract— **Base on innovation resistance theory, this research builds the model of factors affecting consumers' resistance in using online travel in Thailand. Through the questionnaires and the SEM methods, empirical analysis results show that functional barriers are even greater sources of resistance to online travel website than psychological barriers. Online experience and independent travel experience have significantly influenced on consumer innovation resistance. Social influence plays an important role in this research.**

Keywords- innovation resistance theory; online travel; e-tourism; consumer resistance; Thailand

I. INTRODUCTION

In human history, the scientific progress has profound influence on the society. But it also takes much time for people to leave the status quo and adapt to a new innovation. When faced with an innovation, consumers may show adoption or resistance, which are all their responses toward a new technology [1]. Online tourism is an innovative service emerging during the internet evolution. And online consumer behavior, since the birth of online tourism, attracted much attention of scholars, most of which on the adopters and less on the non-adopters and late-adopters. So we only have a flood of literature on the using intention of the online consumers, but few of the non-adopters and the late-adopters. Thus there exists pro-innovation bias [2].

Thailand is one of the countries that tourism industry have grown up largely and become an important traveling destination in ASEAN. In 2014, there were 26.7 million tourists who visited Thailand and income (US$42 billion) ranks number 7 of the world [3]. Although tourism industry in Thailand is expanding and online purchasing becomes more popular, e-commerce is not developed much. Online travel service is hardly used by general public except by few groups of people [4]. Most consumers still purchase products or tourism services with traditional method or offline. Thus studying the causes to find out why those consumers resist the online travel is very important for the tourism entrepreneurs to improve the online travel services and develop e-commerce of Thai tourism in the future.

This research aims to study the factors that made Thai consumers refuse to use online travel websites in purchasing tourism products. Innovation resistance theory was applied in the research. And considering the involvement of consumers, we introduce the factors of consumers' ability, motivation (online experience and independent travel experience) and social influence to frame the model of factors that affect consumers' resistance towards online travel websites.

II. LITERATURE REVIEW AND HYPOTHESIS

A. *Innovation resistance theory, IRT*

Innovation resistance happens when an innovation challenges the satisfactory status quo of consumers or their belief structure [5]. Resistance is a normal consumer response to an innovation and adoption may begin only after initial resistance has been overcome, thus innovation resistance has been regarded as a crucial factor for the success of the adoption of a new technology. Also, the adoption has been described as the result of overcoming resistance [6].

Ram [5] argues that consumer resistance to an innovation is caused by functional and psychological barriers. Functional barrier is an obstacle that has direct impacts on the reluctance of consumers, it can be divided into the usage barrier, value barrier and risk barrier. The psychological barriers are often caused by conflict with consumers' prior beliefs, it including traditional barrier and image barrier.

Since this theory was proposed, it has been applied to the field of e-commerce. In non-adopter group, Laukkanen et al. [7] found that psychological barriers are even greater sources of resistance to internet banking than functional barriers. Lee [8] discovered that the perceived of usefulness and the perceived ease of use are the main factors that affect consumers' using intention and resistance. Tsai [9] indicated that the consumer resistance toward the smartphone also relates to functional barriers and psychological barriers. Lian and Yen [10] found that major barrier for people who refuse to shop online include value and tradition. Based on the innovation resistance theory and related researches this study proposed the following hypotheses:

H1: Usage barrier(UB) have a positive effect onto offline consumer's resistance toward using online travel website.

H2: Risk barrier(RB) have a positive effect onto offline consumer's resistance toward using online travel website.

H3: Value barrier(VB) have a positive effect onto offline consumer's resistance toward using online travel website.

H4: Traditional barrier(TB) have a positive effect onto offline consumer's resistance toward using online travel website.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.42

H5: Image barrier(IB) have a positive effect onto offline consumer's resistance toward using online travel website.

B. Consumer involvement

In the information processing context, Petty and Cacioppo [11] picked out that the consumer involvement, which is largely influenced by ability and motivation, has great impact on how individuals select their ways to process information. So when a consumer reaches a travel website---a platform of information system, their attitudes, whether positive or negative, will be affected by their motivation and ability. On the platform of online travel, online experience and travel experience are the main ability and motivation of consumers and should be taken into the study model as important factors.

1) Online experience, OE: Internet is a key instrument in purchasing online products or services, therefore, the lack of online experience leads to low level of online purchasing ability which is the main ability of online purchasing and consequently brings usage barrier. In the study of consumers using Self-Service Technologies, Kuan [12] finds that when consumers think they have enough personal resources, they perceive less risks and show less innovation resistance. According the literatures, the online experience of consumers probably connects with traditional barriers, because if consumers haven't got acquainted with internet, they don't have the ability and the motivation to use online travel websites, they would insist on their traditional booking habits. Therefore we propose hypotheses as below:

H6: Online experience have a negative on usage barrier of using online travel website.

H7: Online experience have a negative on risk barrier of using online travel website.

H8: Online experience have a negative on traditional barrier of using online travel website.

2) Independent travel experience, ITE: Recently, independent travel style has become popular in the modern travel period. Because independent travel has a distinctive point at freedom, therefore, tourists need to find information for their journey by themselves [13]. Internet has become an important tool for those tourists. This is different from traveling with travel agents and business travel. If they travel with travel agents, the activities will be planned and arranged by the company. The tourists don't need to search for the information by themselves. Jensen [14] found that frequency in traveling has a positive relation with intention in consumers' searching information and purchasing products online. For those offline consumers who hardly go for independent travel, online traveling products or service do little good for them, so their using intention might be faced with value barriers. Based on the literature review, this study proposed the following hypotheses:

H9: Independent travel experience have a negative on value barrier of using online travel website.

C. Social influence, SI

When an individual hasn't the relative ability and enough motivation to process the information deeply, he or she may form their attitude by disposing marginal information and clues towards what they are faced with [11]. People's perceptions of the usefulness of a service or technology may increase in response to persuasive social information [15]. Some researches show that social influence can reduce risk awareness [16]. Chang et al. [17] found that social influence have significantly positive impact on image of Podcasting in tourism. So this research proposes the following hypotheses:

H10: Social influence have a negative on risk barrier of using online travel website.

H11: Social influence have a negative on value barrier of using online travel website.

H12: Social influence have a negative on image barrier of using online travel website.

A representation of conceptual model and all the hypothesized relationships in this study is shown in Fig. 1.

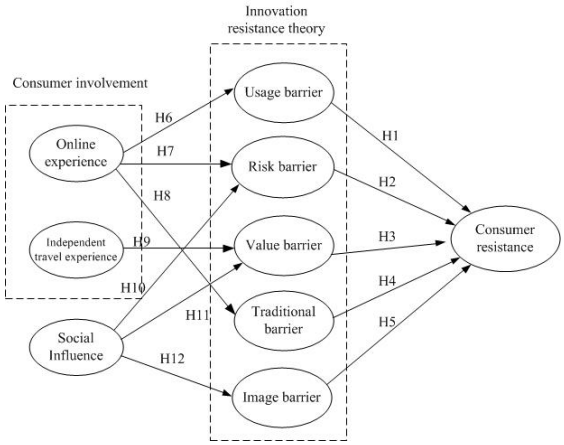


Figure 1. Conceptual model

III. METHODOLOGY

Survey study is employed in this research. The questionnaire was the instrument used in this study to collect the data. Questionnaire contains 9 constructs 26 items. The measurements for each constructs were adapted from many researches as the following; usage barrier, value barrier, risk barrier, traditional barrier and image barrier are adapted from [7], online experience from [18] and independent travel experience from [14], social influence from [15] and consumer resistance from [7] respectively. The population in this research is the internet users who have never purchased travel products online and the samplings are chosen by convenient sampling. The survey uses a field research and online research in two ways. There were 464 copies of questionnaire returned. After eliminating the missing data, there were 415 complete and usable copies of questionnaire which equals a 89% of response rate.

44.3% of the respondents in this study are male and 55.7% are female. Most of them are between 18-25 years old (39.3%). Most respondents hold a Bachelor's degree (57.1%). The first occupations that the samplings do are government/ state enterprise officers (26.5%). The highest income the respondents earn are 15,001-20,000 Baht (35.7%).

TABLE I. THE RESULT OF FACTOR ANALYSIS

Indicators	*Factor loading*	*AVE*	*CR*	*Alpha*
CR1	0.849	0.773	0.911	0.853
CR2	0.913			
CR3	0.874			
IB1	0.890	0.790	0.919	0.867
IB2	0.885			
IB3	0.892			
OE1	0.902	0.816	0.899	0.774
OE2	0.904			
ITE1	0.890	0.776	0.912	0.856
ITE2	0.890			
ITE3	0.853			
RB1	0.848	0.727	0.889	0.813
RB2	0.843			
RB3	0.867			
SI1	0.875	0.765	0.907	0.847
SI2	0.868			
SI3	0.880			
TB1	0.881	0.789	0.918	0.867
TB2	0.887			
TB3	0.898			
UB1	0.873	0.769	0.909	0.850
UB2	0.883			
UB3	0.873			
VB1	0.862	0.764	0.907	0.846
VB2	0.887			
VB3	0.874			

IV. DATA ANALYSIS AND RESULTS

A. Reliability and validity testing

The adequacy of the measurement model was assessed by Smart PLS validity tests. In measuring construct validity, Table 1 shows that all indicators have factor loadings between 0.843-0.913, AVE between 0.727-0.816, and Table 2, all the AVEs are greater than the corresponding squared inter-construct correlations. Cronbach's α, and composite reliability (CR) was used to measure the reliability of the scales. Generally α and CR should be more than 0.7. From Table 1, α=0.774-0.867 and CR =0.889-0.919. According to [19], the measurement model parameter estimates and diagnostics provide strong evidence for the reliability and validity of construct measures.

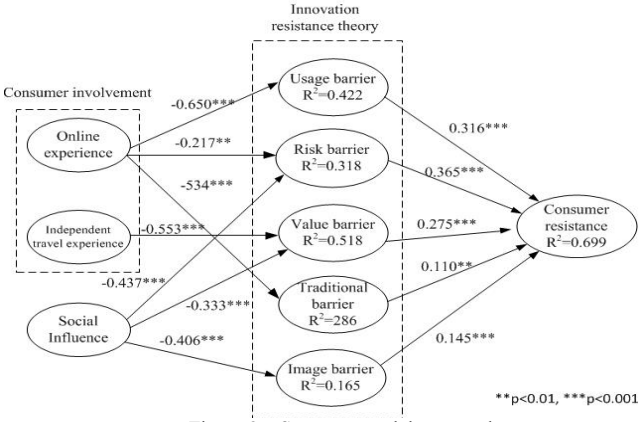


Figure 2 Structure model test result

B. Structure model and hypothesis testing

Fig. 2 displays the results of the structural model test. The findings indicate that the proposed model could explain up to 69.9% of the total variation in consumer resistance. To test whether path coefficients differ significantly from zero, *t* values were calculated using bootstrapping. The non-parametric bootstrapping procedure was applied with 415 cases, 5000 subsamples and no sign changes. The analysis revealed that all of proposed relationships were significant.

Usage barrier (H1: t=7.316, p< 0.001), risk barrier (H2 ： t=8.173, p<0.001), value barrier (H3: t=6.401, p<0.001), traditional barrier (H4: t=2.853, p<0.01), image barrier (H5: t=5.073, p<0.001) was significant effect onto consumer's resistance toward using online travel website, thus, H1-H5 are supported. Online experience have a negative effect on the usage barrier (H6：t=15.378, p<0.001), a negative effect on risk barrier (H7: t=3.892, p<0.01) and a negative effect on traditional barrier (H8: t=14.246, p<0.001) significantly. H6-H8 are supported. Independent travel experience (H9: t=14.363, p<0.001) was significant negative effect on value barrier. Thus, H9 are supported. Lastly, Social influence have significant negative effect on risk barrier (H10: t=7.754, p<0.001), value barrier (H11: t=7.641, p<0.001) and image barrier (H12: t= 9.018, p<0.001) Therefore, H10- H12 are also supported.

TABLE II. INTER-CONSTURCT CORRELATIONS AND THE SQUARE ROOT OF AVE.

Constructs	*CR*	*IB*	*OE*	*ITE*	*RB*	*SI*	*TB*	*UB*	*VB*
CR	**0.879**								
IB	0.410	**0.889**							
OE	-0.529	-0.225	**0.903**						
ITE	-0.325	-0.239	0.020	**0.881**					
RB	0.632	0.235	-0.401	-0.205	**0.853**				
SI	-0.671	-0.406	0.420	0.273	-0.529	**0.875**			
TB	0.504	0.200	-0.534	-0.072	0.341	-0.353	**0.889**		
UB	0.606	0.193	-0.650	-0.095	0.308	-0.457	0.523	**0.877**	
VB	0.591	0.349	-0.202	-0.644	0.356	-0.484	0.272	0.335	**0.874**

Note: Bold diagonal elements are the square root of AVE, which should exceed the off-diagonal inter-construct correlations for adequate discriminant validity.

V. DISCUSSION AND CONCLUSIONS

Previous research has shown that psychological barriers are even greater sources of resistance than functional barriers [7]. The results of this study show that functional barrier has more impact on consumer innovative resistance than psychological barriers. Risk barrier, usage barrier and value barrier is the biggest obstacle for Thai consumers resistance to using online travel websites. The reason might be that functional barrier stays more close to what consumers can perceive.

On the involvement of consumers, the results show that online experience and independent travel experience are the main ability and motivation of consumers. The online experiences negatively influence the usage barrier, traditional barrier, and risk barrier. In other words, if consumers haven't sufficient online experience, they are not able to handle it and perceive lots of difficulties. Then they are stopped by various risks ahead. The results also implicate that independent travel has negative impact on value barrier, their traveling means also influences their views on whether websites is useful. Independent travelers think online travel websites aren't of any use and bump into high value barrier.

When a consumer has low degree of involvement of online travel websites, they would judge the websites in other ways. And according to the research, social influence negatively affects functional barrier and psychological barrier. It has negative impact respectively on the value barrier and risk barrier of functional barrier, and on the image barrier of psychological barrier. So the social influence plays an important role in this research.

At present, Thai online travel is not popular enough, most of the consumers have not used it yet. The result of this research may have some reference effect on the tourist enterprises and related organizations to better their online service and increases online consumers. Of course, there still exist limitations. First of all, this research, on the base of innovation resistance, this research merely considers the consumer involvement and adds the online experience, independent travel experience and social influence to the research model. The subsequent studies can be conducted from other perspectives, like travel enterprises, to reveal deep reasons of consumer resistance to online travel. Secondly, the research is aimed to discuss the factors that influence the using resistance of the consumers, not including those who have used online travel websites. So to get a deeper knowledge of the using intention and behavior of Thai consumers, subsequent studies should be conducted with comparative study among different sorts of consumers.

REFERENCES

[1] V. Cornescu and C.-R. Adam, "The Consumer Resistance Behavior towards Innovation," Procedia Econ. Financ., vol. 6, no. 13, pp. 457–465, 2013,doi: 10.1016/S2212-5671(13)00163-9

[2] S.Ram. " A model of innovation resistance," Advances in Consumer Research ,vol. 14,pp. 208–212, 1987.

[3] World Tourism Organization.UNWTO Tourism highlights 2014 edition.Madrid, Spain:UNWTO,2014.

[4] Jarinee Santijanyaporn, et al. Consumer behavior in e-tourism. Bangkok: Suan Dusit Rajabhat University, 2014, pp. 78-80.

[5] S. Ram and Jagdish N. Sheth, "Consumer Resistance to Innovations: The Marketing Problem and its solutions," J. Consum. Mark., vol. 6, no. 2, pp. 5–14, 1989, doi: 10.1108/EUM0000000002542

[6] I. Szmigin and G. Foxall. Three forms of innovation resistance: the case of retail payment methods. Technovation, vol. 18, no. 6, pp. 459-468, 1998, doi: 10.1016/S0166-4972(98)00030-3

[7] P. Laukkanen, S. Sinkkonen, and T. Laukkanen, "Consumer resistance to internet banking: postponers, opponents and rejectors," Int. J. Bank Mark., vol. 26, pp. 440–455, 2008, doi: 10.1108/ 02652320810902451

[8] S. Lee, "An integrated adoption model for e-books in a mobile environment: Evidence from South Korea," Telemat. Informatics, vol. 30, no. 2, pp. 165–176, 2013, doi: 10.1016/j.tele.2012.01.006

[9] M. Tsai. "Investigating the innovation resistance of smart phone usage in Taiwan," Management of Engineering & Technology (PICMET), 2014 Portland International Conference on. IEEE, 2014, doi: 10.1016/S0166-4972(98)00030-3

[10] J. W. Lian and D. C. Yen, "To buy or not to buy experience goods online: Perspective of innovation adoption barriers," Comput. Human Behav., vol. 29, no. 3, pp. 665–672, 2013, doi: 10.1016/j.chb.2012.10.009

[11] R.E. Petty and J.T. Cacioppo. The elaboration likelihood model of persuasion. In L. Berkowitz (ed.), Advances in Experimental Social Psychology, vol. 19. Amsterdam: Elsevier, 1986, pp. 123–205,doi: 10.1007/978-1-4612-4964-1_1

[12] F-Y. Kuan. , Y-P. Ho, T-F. Chang and C-H. Chen. "E-Commerce Technology Innovation Resistance on Perceived Control and Risk," Advances in information sciences and service sciences. vol. 4, pp. 51–60, 2012, doi: 10.4156/aiss.vol4.issue22.7

[13] K. F. Hyde and R.Lawson. " The nature of independent travel," Journal of Travel Research, vol. 43, pp.13-23, 2003, doi: 10.1177/0047287503253944

[14] J. M. Jensen, "Shopping Orientation and Online Travel Shopping : the Role of Travel Experience," International Journal of Tourism Research, vol. 14, pp. 56–70, 2012, doi: 10.1002/jtr.835

[15] V. Venkatesh and F. D. Davis. "A theoretical extension of the technology acceptance model: Four longitudinal field studies," Management science, vol. 46, pp. 186-204, 2000, doi: 10.1287/mnsc. 46.2.186.11926. Eriksson, "User categories of mobile travel services," J. Hosp. Tour. Technol., vol. 5, pp. 17–30, 2014, doi: http://dx.doi.org/ 10.1108/JHTT-10-2012-0028

[16] H. Kim, T. (Terry) Kim, and S. W. Shin, "Modeling roles of subjective norms and eTrust in customers' acceptance of airline B2C eCommerce websites," Tour. Manag., vol. 30, no. 2, pp. 266–277, Apr. 2009, doi: 10.1016/j.tourman.2008.07.001

[17] S. E. Chang, Y. T. J. Jang, and C. K. R. Chiu, "E-tourism: Understanding users' intention to adopt Podcasting in tourism," Proc. 2012 IEEE Netw. Oper. Manag. Symp. NOMS 2012, pp. 1360–1363, 2012,doi: 10.1109/NOMS.2012.6212074

[18] C. Morosan and M. Jeong. "Users' perception of two types of hotel reservation web sites," International Journal of Hospitality Management, vol 27, pp. 284–292, 2008, doi: 10.1016/j.ijhm.2007. 07.023

[19] C. Fornell and D.F. Larcker. "Evaluating structural equation models with unobservable variables and measurement error," Journal of marketing research, vol. 18, pp. 39-50, 1981 , doi: 10.2307/3151312

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Cloud Data Migration Method Based On PSO Algorithm

Geng Yushui
School of Information
Qilu University Of Technology
Ji Nan, China
gys@qlu.edu.cn

Yuan Jiaheng
School of Information
Qilu University Of Technology
Ji Nan, China
venusyjh@163.com

Abstract—Cloud storage system can play an important role in large-scale, and it supports high-performance cloud applications. To cloud storage systems, data migration is key technology to realize the nodes dynamically extensible and elastic load balancing. How to reduce migration cost of time is the problem that cloud service providers need to solve. Existing research efforts were focused on the data migration issues under the non-virtualized environments , which often do not applicable to cloud storage systems. In response to these challenges, we put data migration issues into the load-balancing scenarios to solve. We propose an algorithm based on particle swarm optimization algorithm which can reduces the cost of time. In the experiment, we can use Yahoo services benchmarking YCSB tool which could verify the validity of the method. It is a test framework designed to help users understand the different cloud computing, database performance.

Keywords-Cloud data; Data migration; PSO algorithm; Load balance

I. INTRODUCTION

A. Background and purpose

Data storage with high reliability and scalability is a huge challenge to Internet companies, traditional databases can hardly meet the demand. In situations like this, cloud store would be a good choice. To storage systems deployed in a cloud environment, data migration is the key techniques to realize nodes dynamically extensible and elastic load balancing. But a lot of sync in data migration process will bring certain impact on system performance, therefore, how to reduce the cost of time is the problem that cloud service providers must solve. However, the status of the storage system, new virtualized environment, strict low latency requirements of consumers as well as the unpredictability and variability of access load have brought new challenges to data migration.

B. Related work

Scholars at the University of Connecticut have conducted researches on how to reduce data migration time of heterogeneous storage system under different transmission capacity constraints[1]. Author takes this as multi-edge coloring problem, considering that how to minimize the time of migrations by optimizing the data migration schedule, that is, using the least number of colors[2]. This work focuses on operations scheduling problems after the institution of migration plan.

Washington State University, Chiu, who studies the cached node data migration problem and proposes a data migration strategy based on greedy method[3].It aims for data migration issues in Key/Value storage system primarily [4]. The basic idea is that, giving priority to transferring part data of hot partition to the neighbor node with lighter loads by using statistical methods partition monitor hot spots. To simplify the migration process, hash algorithms will keep the order of the Key values[5]. The main problems of two parts is that they have not take migrating overhead into consideration.

At the university of California, Santa Barbara, Das and others work mainly for multi-tenancy scenario database cluster data migration issues, the author propose a lightweight, data migration method based on iterative replication, the goal is to minimize the migration overhead[9][10].

C. This paper work

Firstly, we study and analyze the background of data migration. In this paper, data migration issues are solved under the load-balancing scenarios , using a load-balancing framework based on MAPE control loop .We propose a data migration approach that reduce migration time.

When load balancing of cloud storage system become slant, to return to equilibrium, data migration is necessary[6]. Detection of load balancing can be conducted through the entropy value. The higher the entropy value, the more uniform the load distribution is. In this process, you can set a threshold. When the entropy reach this threshold, system performs data migration. To improve the efficiency of the data migration process and reduce resource waste of the system, we use PSO in the data migration process [7].

Innovation of this paper is that we apply PSO algorithm of the swarm intelligence optimization algorithm to a storage system under cloud environment . The algorithm can reduce the cost of time , and improving load balancing degree. Experimental results show that this method can effectively eliminate the load tilt while optimizing the time cost in data migration process.

II. THE PSO MODEL

A. Description of the problem

To more effectively identify hot data and minimize the amount of data migration, the data are divided in the form of partition in each storage node, that is, a data storage node can

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.43

be divided into multiple partitions. Zoning is the basic unit of data migration and load monitoring . Before using the data migration algorithm, we should classify the storage nodes into collection of immigration nodes In_set and Out_set separately according to the relations between normalized loading value and 1/n . In_set contains the node thatload value below 1/n and Out_set contains the nodes that the load value higher than 1/n[8].

Target of Cloud data migration is to achieve a system load balance between nodes in a cloud environment, that is, it should migrate a part of the data of Out_set node to the node of In_set. For such problems, particle swarm optimization algorithm calculate the average shortest time of each data partition migrated between nodes. The algorithm can control the global and local data migration and ultimately it achieve optimize system resources.

The whole process should follow these principles:

*a)*Principle of proximity, the group can perform simple calculation of space or time, the selection of node in In_set is based on the principle of proximity.

*b)*When the system meet the threshold setting of the unbalanced degree, we use PSO algorithm for each node of out_set at the same time.

B. Setting up the PSO model

*a)*Assuming that the entire storage system is N-dimensional to explore, the initial population X is composed of m partitions.

$$X = \left(x_1, x_2, \cdots, x_i, \cdots x_m\right)^T \quad (1)$$

Location of each particle is represented by N-dimensional vector:

$$x_i = \left(x_{i1}, x_{i2}, \cdots x_{iN}\right)^T \quad i=1,2,\ldots m \quad (2)$$

Migration speed of each particle :

$$v_i = \left(v_{i1}, v_{i2}, \cdots v_{iN}\right)^T \quad i=1,2,\ldots m \quad (3)$$

*b)*Calculate the fitness value of all partitions. After finding the optimal solution , the partition can update their speed and position according to the following formula:

$$v_i^{k+1} = v_i^k + c_1 \times rand_1^k \times (Pbest_i^k - x_i^k) + c_2 \times rand_2^k \times (Gbest^k - x_i^k) \quad (4)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (5)$$

c)Parameter Description and Significance

v_i^k, x_i^k is the speed and position of partition i in the k-th iteration;

m: population size, i = 1,2,3 ..., m;

c1 and c2: learning factor or acceleration factor. Separately adjust the maximum step to Pbest and Gbest direction, reflecting the impact of partitions' individual experiences and group experience on data migration trajectory, and reflecting the exchange of information among partitions.

rand1 and rand2: random number between [0,1],increases the randomness of particles flight .

$Pbest_i^k$ is the partition position in the individual extreme points;

$Gbest^k$ is the position of global extreme point of the whole population;

Iteration termination conditions: usually set the maximum number of iterations T_{max}, calculation accuracy ε or maximum stagnation steps of optimal solution△t.

The maximum speed limit vmax: To prevent the partition keeping away from the search space, each dimensional velocity of the particles should in $[-v_{max}, + v_{max}]$, assuming the definition of the search space is the interval $[-x_{max}, + x_{max}]$.

When vi ⩾ v_{max} , vi = x_{max}

When vi ⩽-v_{max} , vi = -x_{max}

If v_{max} is too high, the partition may fly over excellent area. However, if the value is too low, the particles may not fully explore the area while trapped in local optimization area.

*d)*The fitness function

$$F_i(x_1, x_2, x_3) = x_1\text{^}2 + x_2\text{^}2 + x_3\text{^}2 \quad (6)$$

Among them:

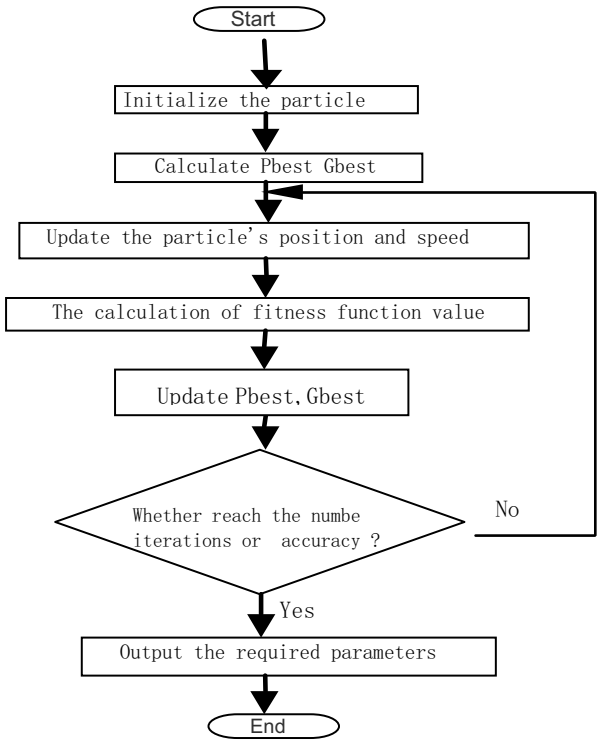
x_1 represents the throughput rate of network, the amount of data transmission on the network per unit of time

x_2 accommodation in the process of the moving nodes

x_3 represents the distance between partition of emigration node and the immigration node.

i represents emigration node.

C. Algorithmic process



Step1: set the algorithm parameters: size of population and dimensions, maximum number of iterations T_{max} or the expected detected entropy , the inertia weight W_{max}, W_{mix} and the optimal solution set.

Step2: Initialize each population, calculate the fitness value of each particle F_{ij} (x_1, x_2, x_3).

Step3: Particles conduct migration in the system in accordance with each particle's value F_i (x_1, x_2, x_3) in which following the principle: follow the migration process, the particles move to immigration node with the minimal cost (F_i (x_1, x_2, x_3) value is the smallest node) .The fimess value of each particle compare with Pbest, if the current position is better, it will be the best position Pbest;

Step4: For each particle, compare its fitness value and the best position it passed Pbest .If the former is better, take it as current best position.

Step5: Obtain particle's new speed and position based on the formula of iterations . When the load value of immigration node is equal to the total load value /number of nodes of the threshold value of system, then the node is automatically removed.

Step6: the termination condition: the number of iterations reaches the maximum number of iterations set or achieve set detection value (entropy) of load balancing

D. Pseudo code of PSO

```
// Function: pseudo-code of PSO
// Note: this example problem aims for minimum fitness for
   the purpose
// Parameters: N as the population size, that is,the number of
   partitions
Procedure PSO
    For each particle i
     Initialize velocity Vi and position Xi for particle i
          Evaluate particle i pBest and gBest
End for
gBest=min {pBest}
while not stop
for I to N
      Update the velocity and position of particle i
      Evaluate particle i
      If fit(Xi)<fit(pBest)
         pBest=Xi;
      If fit(pBest)<fit(gBest)
         gBest=pBest;
end for
end while
print gBest
end procedur
```

III. EXPERIMENTAL ANALYSIS

A. Experimental environment and Settings

Experiment environment consists of 4 identical blade servers (2Intel Xeon Quad-Core E5620 2.4GHz,16G memory), virtualization management software Citrix Xen-server 6.0, We use the ElastiCamel cloud storage system developed by Chinese Academy of Sciences Institute of Software. The system consists of 30 data storage nodes and one managed node, the Key/Value storage node will be responsible for data services and data migration implementation, management node is responsible for membership management, load balancing, routing table maintenance testing and migration planning and so on. Data routing adopt client routing mechanism.

For the data partition, we use the improved consistent hashing algorithm proposed by Amazon, which introduce virtual nodes[8]. The size of the hash space is divided into several equal Q data partition (also known as virtual nodes, Q>>N,N is the number of nodes, and Q is the number of partitions), each node assign a different number of data partitions according to its handling ability. In the concrete implementation, the number of storage node is 30, the total number of partitions is set to 512, therefore, each storage node assigns about 16~18 partition.

B. Detection data and image of imbalance

The main objective in this section is to evaluate the cost of migration time . Load model is applied in this experiment to simulate the load tilt scene. Management node is responsible for sending data migration plan to the relevant storage node. To avoid blocking network I/O, multiple data migration operation adopt serial execution. Concept of non-equilibrium degree is also introduced in this experiment. When the non-equilibrium degree is above a certain threshold, re-equalization operation would be triggered. Here, the non-equilibrium threshold is 0.05 based on experience. Greedy algorithm and particle swarm optimization algorithm are used to optimize load balancing, and each optimization algorithm runs 10 minutes. Finally, we use the data of experiment to draw images of non-equilibrium systems.

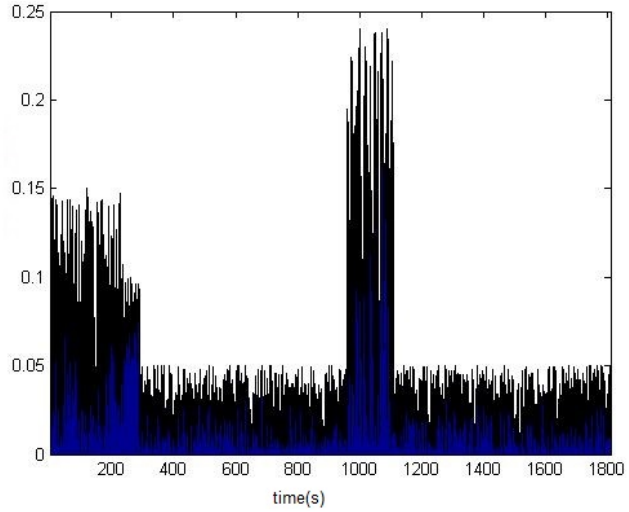


Figure 1. The image of the greedy algorithm

Figure 1 shows the changes of the non-equilibrium degree in the whole system .We can found that when the load is non-equilibrium, the greedy algorithm is 240 seconds

in system's initial running .After that, the system keeps the load balancing state until the 900 second, then the new load mode is activated, the system is under load non-equilibrium state again. In the 960 second, the system return to the state of load balancing once again.

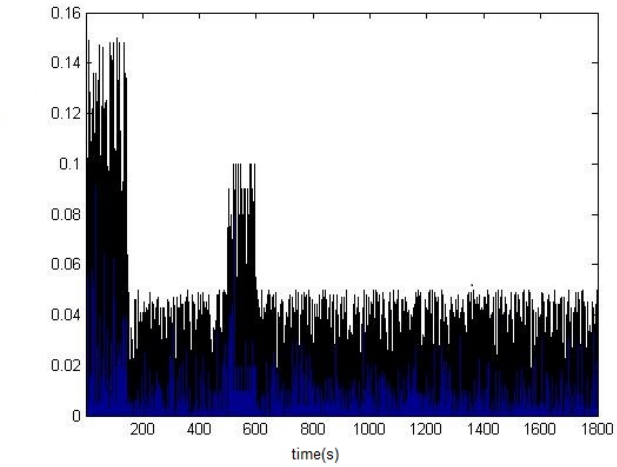


Figure 2. The image of particle swarm optimization algorithm

Figure 2 also shows the changes of the non-equilibrium degree in the whole system. We can found that when the load is non-equilibrium, the PSO algorithm is 180 seconds in system's initial running. After that, the system keeps the load balancing state until the 500 seconds, then the new load mode is activated, the system is under load non-equilibrium state again. In the first 630s, the system return to the state of load balancing once again.

In summary, compared to the greedy algorithm, the PSO algorithm is applied in data migration process. The experimental results show that the algorithm can reduce the cost of time. The proposed algorithm can be faster to achieve the required load distribution uniformity. The experiment achieved the desired effect.

IV. THE DEFICIENCY OF THIS ARTICLE

In the experiment, in order to reduce the complexity of the problem, we assume that all storage nodes are configured with the same prediction model, but the experiment does not take into account the node CPU and memory heterogeneous. In the future, we will study the heterogeneous nodes which have an impact on the method.

This method is not applicable to the system which produce sudden load continuously. Therefore, if the stability period is lower than the data migration time, the execution migration will introduce unnecessary expenses. The next work intends to consider the load stable period, we can use time series model and cost model of the system, and make improvement on the quality of the data migration strategy in the further.

ACKNOWLEDGEMENT

[1]Project supported by the projects of Shandong Province Higher Educational Science and Technology Program,China (No. J12LN20).

[2]Project supported by the projects of Shandong Province Science and Technology Development Plan,China (No. 2014 GGX101052).

[3]Project supported by the projects of Shandong special independent innovation and achievements transformation,China (No. 2014ZZCX03408).

[4]Project supported by the projects of SShandong province natural science foundation,China (No. ZR2014FQ021).

REFERENCES:

[1] Kunkle D, Schindler J. A lsoad balancing framework for clustered storage systems.,n:Proc.of the 15th Int'l Conf. on High Performance Computing (HiPC 2008). 2008. 57-72.

[2] Wang Hao, Summary of graph coloring problem,ELECTRONIC TECHNOLOGY & SOFTWARE ENGINEERING, 2014.8.

[3] Chiu D, Shetty A, Agrawal G. Elastic cloud caches for accelerating service-oriented computations. In: Proc. of the ACM/IEEE Int'lConf. for High Performance Computing, Networking, Storage and Analysis (SC 2010). 2010. 1-11.

[4] Pfaffhauser F. Scaling a cloud storage system autonomously [MS. Thesis]. Zuerich: Eidgenössische Technische Hochschule Zürich, 2010.

[5] Huang Qiulan;Cheng Yaodong;Chen Gang,Computing Center, Institute of High Energy Physics,Chinese Academy of Sciences; 2014.01.

[6] Mei Y,Liu L,Pu X,Sivathanu S,Dong X,Performance analysis of network I/O workloads in virtualized data centers, IEEE Trans.on Service Computing,2011.

[7] Duan Huyi,Li Gang,Ru Hanrong,Chen Xin. Database evaluation In virtualization environment .Journal of Air Force Early Warning Academy, 2013.6.

[8] Qin Xiulei,Zhang Wenbo,Wang Wei,Wei Jun,Zhao Xin,Zhong Hua,Huang Tao,Data migration based on method Sensitive overhead for cloud Key/Value storage system.Journal of Software, 2013,24(6).

[9] Das S, Nishimura S, Agrawal D, Abbadi AE. Albatross, Lightweight elasticity in shared storage databases for the cloud using live data migration. In: Proc. of the 37th Int'l Conf. on Very Large Data Bases (VLDB 2011). 2011. 494-505.

[10] Zhuang Jin,Kou Wei,Li Peng. Autonomy of cloud services model in the dynamic design of data migration [J]. Computer Development & Applications. 2011(09)

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Virtual Machine Migration Strategy in Cloud Computing

S. Liyanage
School of Computing and Information Systems
Kingston University
Kingston upon Thames, United Kingdom

S.Khaddaj, J. Francik
School of Computing and Information Systems
Kingston University
Kingston upon Thames, United Kingdom

Abstract— **In a typical cloud based data centre, several physical machines (PM) host dozens of virtual machines (VM), which run various applications and services. VM load varies according to the different type of users' applications and traffic, and sometime this traffic may overwhelm VM's resources. Physical machines that host many VMs have limited capacity and if they are overloaded their performance might degrade or completely fail. A VM machine that hosts several different applications on a heavily loaded PM can be migrated to another underloaded PM in order to exploit the availability of the resources and to balance the load. In data centres, VMs exposed to huge amount of user traffic, if VM migration occurs on a network link that is overwhelmed with user traffic, then it will create a bottleneck if there's not enough bandwidth on the link to support the VM migration. Therefore, there is a clear need for mechanisms to control user traffic on both source and destination PMs in order to provide guaranteed bandwidth that needed for VM migration.**

Keywords- Virtual Machine (VM), Physical Machine (PM), VM Migration, Load Balancing, Cloud Computing

I. INTRODUCTION

Cloud computing has revolutionized the IT industry in recent years. Cloud Computing is a paradigm where processing, storage, and network capacity are made available to users in an on demand manner through virtualisation on a shared physical infrastructure. The Cloud Computing concepts are based on distributed, parallel and grid computing coupled with virtualisation [10].

There are three basic service modules in the Cloud Computing, Software as Service (SaaS), Platform as a Service (Paas), and Infrastructure as a Service (IaaS). Today many organisations are experiencing the benefits of Cloud Computing, they build out private clouds using various tools such as VMware or OpenStack, and establish online services that are not limited to internal users, but outside their firewalls as well.

Cloud computing provides a number of large computing infrastructures for large scale data centers, which contain dozen of physical nodes with multiple virtual machines running on them. These VMs could also be migrated across different physical nodes on demand to achieve various goals. Modern day cloud based datacenters contain hundreds of virtual servers that host different critical applications ranging from those that run for a few seconds to those that run for longer periods of time, some popular cloud based applications are MS Office 365, dropbox, CRM software Salesforce.com, meanwhile cloud based Netflix stream movies and TV programs to members across the world [4].

Virtualisation is the key enabling technology of Cloud Computing that allows simultaneous execution of diverse tasks over a shared hardware platform. Virtual Machine (VM) is a software implementation of a computing environment in which an operating system or program can be installed and run [5]. In Cloud Computing, applications and services are hosted on Virtual Machines that span over several physical servers with dedicated resources (CPU cores, RAM, Disk Space, etc) are allocated to these VM in order to closely match the applications needs. Virtualisation provides many benefits, such as resource utilization, portability, application isolation, reliability, higher performance, improved manageability and fault tolerance [5].

Live migration is a very important feature of virtualisation where a running VM is seamlessly moved between different physical hosts. Source VM's CPU state, storage, memory and network resources can be completely moved to a target host without disrupting the client or running applications.

However, live VM migration can consume significant bandwidth (500 Mb/s for 10 seconds for a trivial web server VM), so these non-negotiable overheads need to be considered when scheduling migration [1]. Higher workload density in combination with network bandwidth intensive migration can lead to network congestion. To secure the bandwidth for VM migration, it's necessary to control user traffic, if we can find a mechanism to predict bandwidth that required for VM migration on the source Physical Machine (PM), and then limiting the user traffic on both source and destination PM, it can guarantee the minimum bandwidth that required to execute VM migration and will avoid network bottlenecks. Given the above analysis, the main objective of this research is to investigate and design a decentralized bandwidth aware autonomic intelligence framework for live VM migration to manage the workload of physical servers.

This paper starts with a brief discussion of the background of virtual machine migration techniques, particularly in terms of memory and disk migration and the impact on physical machine load and bandwidth. This is followed by a presentation of the main architecture of the proposed framework. Finally, a summary of the finding is presented.

II. VM MIGRATION TECHNIQUES

Cloud computing based on virtualisation and utility computing concepts, virtualisation enables multiple and secure virtual servers to run on a single physical server. Virtualisation technology was implemented on IBM mainframe in 1960, and it is the key concept behind cloud computing. In virtualisation, the Virtual Machine Monitor (VMM) also called Hypervisor, which is a software layer that provides resources to emulate a hardware interface for VM to run on. VMM runs on bare hardware or on top of an operating system [6]. There are two types of hypervisors, type 1 and type 2,

Type 1 hypervisors run directly on machine's hardware with VM resources provided by the hypervisor. Type 2 hypervisors run on a host operating system to provide virtualisation services. In this research we only focus on type 1 hypervisors, e.g. VMware ESXi and CitrixXenServer [9].

Live VM migration is a technique that migrate the entire system of a VM, including OS, memory, storage, process and network resources

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.44

and also its associated applications from one physical machine to another without disrupting the client or running applications [7].

Disk State Migration: Disk state migration of VM is an important factor of VM live migration; it involves transferring the VM's virtual hard disk from source to the destination host. Virtual machine consists of one or several virtual hard disks; VM stores its operating system, programs and other data files on its virtual hard disks, so these virtual hard disks need also to migrate when live VM migration occurs. There are two main live storage migration techniques that migrate VM disk images without service interruption. The techniques are, Dirty Block Tracking (DBT) and IO Monitoring, DBT is the widely adopted technique by many VM vendors (e.g. Xen and VMware ESX), it uses bitmap to track write request while the VM image is been copied. Once the entire image is being copied to the destination, a merge process is initiated to patch all the dirty blocks from the original image to the new. The disadvantage of this technique is if number of dirty blocks are not converged due to heavy write request, the migration of VM would be significantly long [11].

The following factors must be considered when performing live VM migration, which are memory, disk and network resources. There are two main methods of migrating a VM, offline migration (cold VM migration) and online/live migration (hot VM migration), in offline VM migration, the services running on VMs are completely stopped during the migration process while live migration method keep all services running on VMs .

Migration Techniques

There are three main categories of migration techniques as follows, [7]:

- Energy Efficient Migration
- Load Balancing
- Fault Tolerance Migration

The main idea behind Load Balancing VM technique is to distribute load across all physical servers in order to avoid bottlenecks, improve availability and over or under provisioning of resources (figure 1.). In Energy Efficient VM Migration energy is conserved by migrating VMs from low utilised servers to servers that have enough capacity to host them, and then the low utilised servers can be shutdown to save energy in the data center. Fault Tolerance VM Migration aims to predict a physical machine failure beforehand and migrate all VMs from failing physical server to another physical server, this technique improves availability in cloud based data centers. However, no matter what algorithms or techniques are used for the migration process, it will involve VM memory migration, Disk migration etc.

VM Memory Migrations: In live VM migration, the most important phase is transferring source VM memory state to the destination VM. Memory migration can be divided into three phases, push phase, stop-and-copy phase, and pull phase. According to Botero [3] in push phase, the source continues running while certain pages are pushed across the network to the new destination. In Stop-and-copy phase, the source VM stopped, pages are copied across to the destination, and then the new VM started. In pull phase, the new VM starts its execution and if it access a page that has not yet been copied, then the page is copied across to the destination. Most migration strategies select either one or two of the above phases, pre copy approach combines push with stop-and-copy phase while post-copy approach combines pull copy with stop-and-copy [3]. To help understand the VM memory migration, VM memory is categorised into five major categories [8].

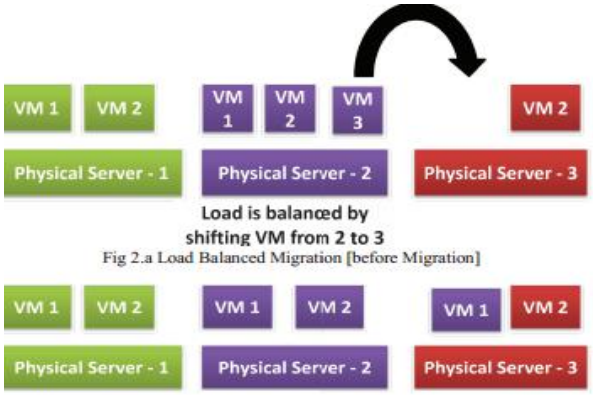


Figure 1 Load balancing Migration [7]

III. The Proposed Framework

The proposed Bandwidth Aware Dynamic Virtual Machine (VM) Migration Framework is aiming to provide two services, working as a load balancer, and facilitating required bandwidth for live VM migration by controlling user traffic dynamically on Physical Machines (PM) at peak times.

Live VM migration consumed significant amount of bandwidth, generally PMs in large cloud data centers are connected to networks that have higher bandwidth, but some applications which are running on PMs may experience huge user traffic during peak times, if VM migration occurs during those peak times, VM migration and user traffic will compete for network bandwidth, then datacentre's network may not have enough resources to support both VM migration and demands of application users, which would create a bottleneck in the network. The proposed framework will dynamically control user traffic on busy physical servers and facilitate the minimum required bandwidth for VM migration during peak times, it will schedule VM migration in an efficient manner.

VM live migration can be helped to provide seamless connectivity and minimal downtime for users. The proposed framework will also work as a load balancer. Physical servers in cloud data centers have limited resources and it may vary according to the workload of the particular PM, some PMs may get huge amount of workload while some get very low user traffic. The proposed framework will have a complete picture of all PMs' available resources and workload, it will dynamically trigger VM migration from overloaded machines to underloaded machines to balance the load of physical machines. The dynamic load balancing strategy will reduce the differences among all nodes by migrating VMs from heavily loaded PMs to lightly loaded machines.

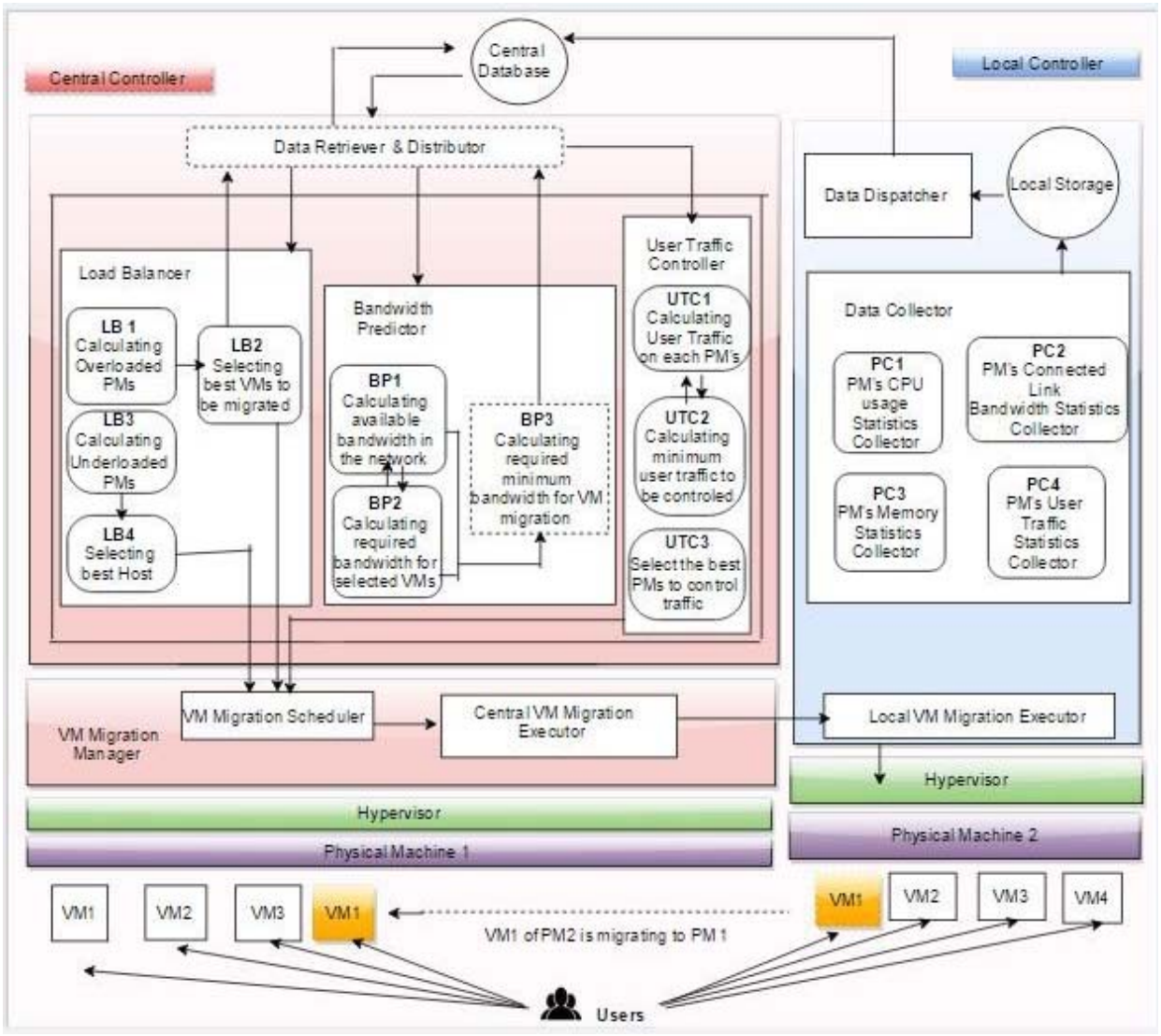


Figure 2: The proposed Framework

The proposed framework architecture is equipped with two main components, Central Controller and Local Controller. The Central Controller deployed on the controller physical node along with the local controller. There is only one central controller for a cluster. On the other hand the Data Collector & Distributor is responsible for fetching data from the central database and distributes it to the relevant components in the framework, and also saving central controllers data on the central database.

The load balancer is responsible for balancing the load on physical machines in the data center, it resides in the Central Controller and it's responsible for making following decisions:

- Detection of overloaded PMs
- Detection of underloaded PMs
- Selection of best VMs to migrate
- Selection of best hosts for migrating VMs

Each PM periodically executes the overloaded detection algorithm to find out overloaded PMs in the datacentre, the algorithm will be based on setting a static CPU and memory utilization threshold to detect overloaded PMs. When the algorithm invoked, it compares the current CPU and memory utilisation of the host PM with the defined threshold. Algorithm detects an overloaded PM if the utilisation of PMs' resources (CPU usage and Memory usage) is exceeded over 90 % of the total PM resources. The PMs underloaded algorithm works in a similar way to the overloaded PM algorithm. It detects an underloaded PM by searching for a host for VM migration and it chooses the first available host that has total resources threshold under 50 % of

PM's total resources. If algorithm can't find a host that has not reached 50 % resources threshold, then the algorithm finds a host with minimal utilisation of resources compared with other PMs in the cluster.

Once the overloaded detection algorithm finds an overloaded host, it invoked an algorithm to find best VMs to offload from the host. An example is using the Minimum Migration Time Policy (MMTP) algorithm to find the best VM/VMs to offload from the host [2]. Once it selected the VMs that need to be migrated, it saves a copy of the list in the database and send a request to the VM Migration Scheduler. The rest of the framework components are described below.

Bandwidth Predictor: It is responsible for calculating available total bandwidth in the network and minimum bandwidth that need for VM migration.

User Traffic Controller: It is responsible for controlling user traffic on the network in order to facilitate required minimum bandwidth for VM migration. This component does following calculations:

- Calculating user traffic on each PM
- Calculating the minimum amount of user traffic to be controlled
- Select the best PMs to control user traffic.

VM Migration Scheduler: This component is responsible for scheduling VM migration and accepting scheduling request from Load balancer, Bandwidth Predictor and User Traffic Controller, it scheduled the requests periodically and pass execution commands to the Central VM Executor.

Central VM Migration Executor: This component acts like the agent between local VM migration executer and VM migration scheduler. It accepts VM migration execution commands from the VM Migration Scheduler and pass them to the Local VM Migration Executor.

Local Controller: The local controller deployed on every PM in cluster. The local controller is equipped with following components.

Data Collector: This component periodically collects, CPU and memory utilization, bandwidth on connected link, and user traffic on each PM, and it also collects the CPU utilization by the hypervisor. Collected data will be saved in the local file based storage.

Data Dispatcher: The dispatcher acts as the central hub that manages all data flow in the Local Controller. It has a simple mechanism for transferring data from the local file based data storage and to the central database.

Local VM Migration Executor: This component is responsible for passing VM migration execution commands to the local Hypervisor. It gets VM migration execution commands from the Central VM Migration Executor.

Data Storages: There are two main data storage systems, central database, deployed on the controller host storing historical data of every PM, and the local file based data storage storing resources usage and other statistics that collected from each PM, the data collector periodically transfer data to the central database.

IV. CONCLUSION AND FUTURE WORK

In this work a framework for the management of virtual machines migration and allocation to PM that is based on resource usage and network bandwidth is proposed. At the core of the framework is the load balancer and the bandwidth monitor and predictor. It is a comprehensive framework for virtual machine migration in a cloud environment, particularly with heterogonous physical machines. The proposed framework and the underlying algorithms can be modelled and simulated in environment such a CloudSim which will be considered in future work.

REFERENCES

[1] Alexandar.S, Setzer.Z.Network Aware Migration Control and Scheduling of Differentiated Virtual Machine Workloads.Technicle University of Munich. 2009

[2] Beloglazov.A. “Energy – Efficient Management of Virtual Machines in Data Centers for Cloud Computing.2013. University of Melbourne.

[3] [3] Botero.D.P.A “ Brief Tutorial on Live Virtual Machine Migration from Security Prespective”.2012.Prenceton University .New Jersy.

[4] Forbes.com.20 Most Popular Cloud Based Apps Downloaded in to Entrprises.Mckendrick.J.Available at:http://www.forbes.com/sites/joemckendrick/2013/03/27/20-most-popular-cloud-based-apps-downloaded-into-enterprises/ [Accessed 26.06.2014]

[5] [5Kapil.D, Emmanuel S, Plli, Ramesh.C.J.”Live Virtual Machine Migration Techniques:Survey and Research Challenges. 3rd IEEE International Advanced Computing Conferece, pp 363-368.2012.

[6] King.S.T,George.W,Dunlap.W, Chen.P.M.” Operating System Support for Virtual Machines.Proceeding of the 2003 USENIX Technical Conference. 2003

[7] Leelipushpam.P.G.J , Shermila.J.”Live Migration Techniques in Cloud Environment: Survey”. IEEE Conference on Information and Communication Technology. pp 408-413.2013

[8] Hu.W, Hicks.A,Zhang.L, Dow.E.M, Sony.V , Jiang.H, Bull.R, Jeanna.N, N.Matthews. “A Quantity Study of VirtualMachine Live Migration”. 2013.Clarkson University .New York.

[9] Virtualizationreview.com.2009.Type 1 and Type 2 Hypervisors Explained. at: http://virtualizationreview.com/blogs/Every day-virtualization/2009/06/type-1 -and-type-2-hypervisors-explained.aspx [Accessed 27.02.2015]

[10] Zhang.Q, Cheng.L,Boutaba.R.”Cloud Computing: State of the Art and Research Challenges. Brazilian Computer Society.2010. pp 7-18

[11] Zhou.R , Liu.F , Li.C , Li.T .”Optimizing Virtual Machine Live Storage Migration in Heterogeneous Storage Environment”. 2013.Ninth Annual International Conference on Virtual Execution Environments .Texas.U.S.A.

A VDI system based on CloudStack and Active Directory

Wei Wei, Yousong Zhang, Yongquan Lu, Pengdong Gao, Kaihui Mu
Communication University of China
Beijing, China
e-mail: 834236156@qq.com, wwwzys@163.com, yqlu@cuc.edu.cn

***Abstract*—Cloud computing is one of the hottest topics in recent years, from the initial Iaas (Infrastructure as a Service) to Paas (Platform as a Service) to Saas (software as a service), the scope of cloud computing continue to expand the bottom up, bringing people's lives profound changes. Compared with the ordinary desktop, cloud desktop has the advantages of high security, high flexibility, high resource utilization, low operation cost and so on. With the development of virtual technology, the combination of remote desktop and virtual operating system makes the cloud desktop more mature. Design a cloud desktop management system, to achieve unified management and maintenance of cloud desktops is of great significance. This paper first discusses cloud desktop and CloudStack, elaborates the technology in cloud desktop deployment, and finally put forward to achieve the system. The main research work are as follows: Firstly, this paper analyzes the deficiencies of CloudStack as a cloud desktop management system. And then this paper completed the overall system architecture design. The system architecture is based on PHP CI framework based on CloudStack cloud platform for dynamic resource allocation. Secondly, this paper discusses the key technologies about the cloud desktop system. Thirdly, this paper carried out the detailed design of the system. Lastly, this paper tested the system prototype.**

Keywords-component; cloud desktop; CloudStack; cloud storage; Active Directory

I. INTRODUCTION

Apache CloudStack [1] is open source software designed to deploy and manage large networks of virtual machines, as a highly available, highly scalable Infrastructure as a Service (IaaS) cloud computing platform.

However, CloudStack support for the desktop is relatively weak, user or administrator can access to the virtual machine through CloudStack CPVM virtual machine (console) VNC [2], data cannot be transferred on VNC Access Console. And VNC cannot adjust the window size, what's worse; VNC has higher delay, so the function is used to manage virtual machines, unable to provide cloud Desktop Services to the user. So it is very necessary to develop a management system of cloud desktop based on CloudStack.

Users can upload files to the cloud storage from the local; users can also download the file to the local. However, if we want to edit and process the files in the cloud, we can only download the file to the local. And upload it to the cloud disk when the file was handled locally. If we could combine cloud storage with cloud desktop, users can process directly documents and data in cloud desktop.it will greatly improve the work efficiency. At the same time, the user could save user's data to the cloud disk when user use cloud desktop.

Cloud desktop and cloud storage management system based on CloudStack and Active Directory. And the system would provide users with a unified service by cloud desktop management platform. The entire system uses the basic networking environment of CloudStack for dynamic resource allocation and use AD to achieve user unified registration, unified authentication service and access control. The system achieved the combination of management and use for cloud desktop and cloud storage.

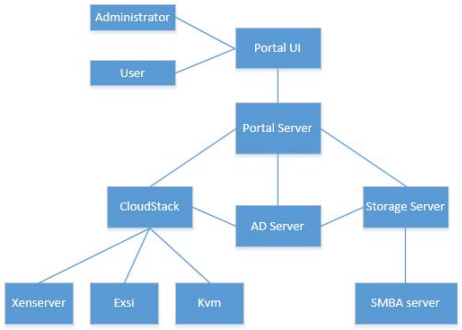


Figure 1: System Architecture

Portal UI is cloud desktop management Interface and user could manage cloud desktop virtual machines by UI interface. The system provides the web-based UI that can be used by both administrators and end users.

Portal server is the server of cloud desktop management system.

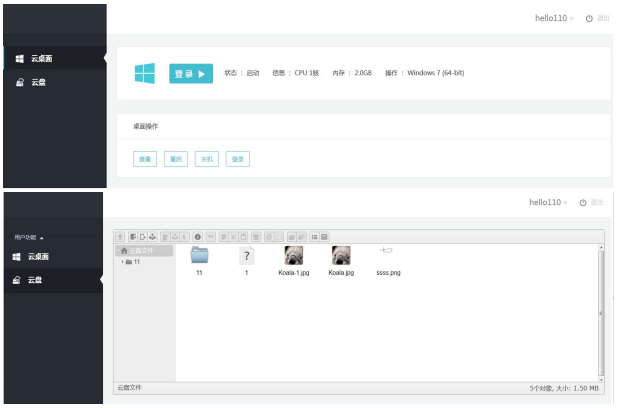


Figure 2: Portal UI

Ordinary users:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.45

- Register function, logon, logoff, and other functions.
- Cloud desktop application, cloud desktop login
- Cloud storage management functions: upload and download files, copy, delete, paste and other related documents basic operating functions

Portal server provides administrators with complete control over the lifecycle of all guest cloud desktops executing in the cloud platform.

Administrator:

- Manage all users
- Manage all cloud desktops
- Manage all cloud storage space
- Configuration CloudStack

AD is Microsoft Active Directory, Active Directory is Directory Services for Windows Standard Server, Windows Enterprise Server and Directory Services Windows Datacenter Server [3]. The primary function of the Active Directory is the client's security management and client's standardized management. The AD domain is one of the most important concepts about Microsoft Network. The AD domain actually refers to a group of servers and workstations, and they agreed to name and password for the user and machine accounts centralized in a shared internal database. The main function is to accept the AD provides user registration, authenticate user logins, cloud desktop virtual machine included AD management, user access restrictions cloud desktops, cloud storage server integrated into AD management, control user access to storage.

CloudStack is used by a number of service providers to offer public cloud services, and by many companies to provide an on-premises (private) cloud offering, or as part of a hybrid cloud solution [4].

CloudStack architecture shown in the illustration, CloudStack mainly consists of the following components, CloudStack Kernel module as the core modules, including virtual machine management, storage management, network management, template management, snapshot management, is mainly for distribution and recycling of all types of resources. Account module deals with user accounts, to ensure that users have the appropriate permissions to access the corresponding resource. Business Logic module for unified management and unified allocation of resources, including policy management, upgrade management, HA Manager and other related services, the user's resource request targeted to lower utilization place. CloudStack Fundamentals include Agent Manager, cluster management, database access layer management. Meanwhile CloudStack provide external interfaces, the resources can be accessed by calling its API, this system is to call on the basis of its API be built.

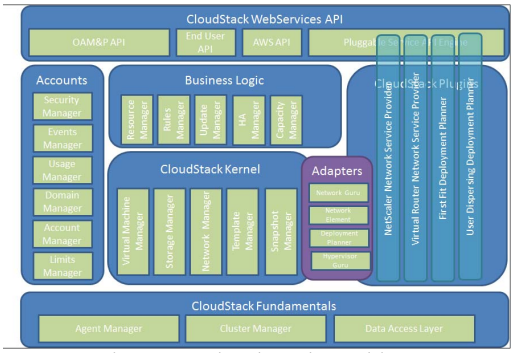


Figure 3: CloudStack Architecture

Cloud storage is a new concept which developed on cloud computing. And it is a new network storage technology. Cloud storage is a service model in which data is maintained, managed and backed up remotely and made available to users over a network (typically the Internet)[5].

Cloud storage of the system mainly provides storage function and cloud desktop storage disk mounting function for the user. The system will allocate storage space for users and users' cloud disk can be displayed on the portal management interface when user registration is successful, and cloud disk will mount to the user's cloud desktop in the form of network storage.

II. RELATED WORK

A. User unified registration, unified authentication mechanism

Users log in successfully and complete users' registration through the portal interface, portal server receives users' request information, after the verification is correct, the users' information were stored in the portal server database, and then the server would send requests to AD server for creating user information. And the storage server would create storage space for the users.AD would export the users' information to CloudStack, CloudStack would create a virtual machine under the user account, complete the user registration function.

CloudStack and AD combined [6]: To set up LDAP authentication in CloudStack, call the CloudStack API command addLdapConfiguration and provide Hostname or IP address and listening port of the LDAP [7] server.

The following global configurations should also be configured.

Storage Server and AD combined: deployed and configured samba server. Samba services including installation, configuration of smb.conf file, open the corresponding port. Samba storage server joins AD domains.

Including import domain certificate operations and join AD domain.

B. Users create cloud desktop

Users can specify a template to create the corresponding cloud desktop, and users can select the corresponding serviceoffering. The portal server would submit virtual machine application program to the CloudStack server by calling CloudStack API. CloudStack would create the virtual

machine, when the virtual machine is created; the cloud desktop is displayed on the user's portal interface.

Creating the cloud desktop is the process that the portal server calls the CloudStack interface to create a virtual machine; the algorithm flow chart, as shown below:

a) Users submit the request of creating cloud desktop on the UI portal interface, and then the request submitted to CloudStack by CloudStack API.

b) CloudStack would check if the User has permission to create a virtual machine

c) According to the user's cloud desktop parameters, check template, serviceoffering, diskoffering, storage, to see whether existing resources can meet the application needs.

d) Check current user's resources.

e) Check the format of the template, if it is ISO format, to create a virtual machine by the ISO mirror, if not, to create a virtual machine by the template.

f) Select the network for the virtual machine and assign the network card information for the virtual machine

g) Allocate UUID for the virtual machine, and make sure that each virtual machine has a unique identifier.

h) Check the name of the virtual machine and other VM in the system conflict, to ensure the uniqueness of hostname

i) Create virtual machine instance.

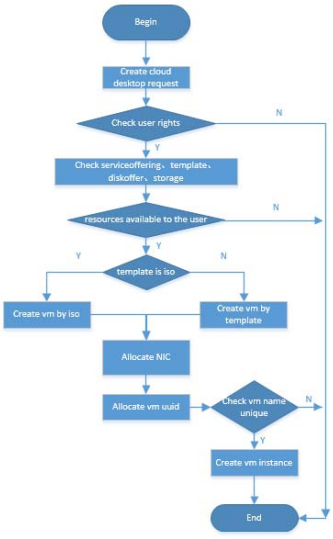


Figure 4: The algorithm flow chart

At the same time, portal also needs to record cloud desktop information in his own database. Portal server would record cloud desktop information which come from CloudStack into the database. These information mainly include cloud desktop IP, name, owner, CPU and memory, MAC address, virtual machine in the status When Cloud Desktop is created.

C. *The management and use of cloud desktop*

a) Manage cloud desktop

User manage cloud desktop by the UI portal interface, including the launch of the cloud desktop, stop the cloud desktop, delete the cloud desktop, restart the cloud desktop, etc.

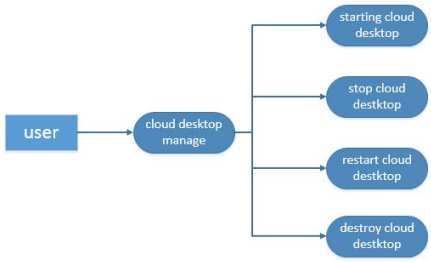


Figure 5: User manage cloud desktop

b) Use the cloud desktop

When the user clicks the virtual machine enable button on the portal UI, the user will automatically enter the management interface of the cloud desktop. Users do not need to enter the user login interface, landing and user authentication automatically completed. The virtual machine which belongs to the user would allow the user login, otherwise, users can not log virtual machine.

c) Implementation process

Any user in the domain can access virtual machines in the domain, how to achieve the function that the cloud desktop can be accessed only by the cloud desktop of a single account.

The cloud desktop template should be created in special way. Including virtual machine in the domain, opening the users Domain function in the virtual machine, changed virtual machine starting strategy.

The template for the generation of the cloud desktop is implemented. When the cloud desktop is started at the first time, the starting script will send request to the portal for querying the MAC address [8] of the current user's desktop cloud from the portal records, and then user information from the portal would return back to the cloud desktop, desktop cloud will write the user's information to configuration files.

When users log in the operating system by RDP [9] (Remote Desktop Protocol), the login script matches the information of the current user with the information of the configuration file, and if it is consistent, then allows login, otherwise it will refuse the user login.

Figure 6: User access cloud desktop

The whole process from the creation of the desktop cloud, cloud desktop joined to a domain, user login, verify user information. Completing the whole process automatically, users only need to click the login button, and no need to enter any information again, the cloud desktop can complete the user of the verification, the login process, the user can easily use of the virtual machine.

D. *The use of the cloud storage*

After the user registration is completed, the system will assign a user a fixed size storage space, users can perform a series of operations on the storage space, including: upload and download files, delete cut copy files, and so on; at the same time, users log in to the virtual the machine, the user's cloud disk space will be in the form of network drives automatically mount to the virtual machine under the user, and cloud disk space and network disk data synchronization, regardless of where the file is created or processed information can be synchronized to another side.

Cloud storage server was added to the AD domain, registered in the AD users on the storage server is visible. In turn, using AD in access to each user's storage directory for each user set when another server or cloud desktop virtual machine to mount the user store that is visible only to the user directory.

Cloud Desktop for cloud storage support: When users log in to the cloud desktop, cloud storage will be in the form of network storage cloud directly mounted onto the desktop, works as follows:

After a successful user login cloud desktops, through logon scripts, to portal administration page sends a request to obtain the user's storage shared directory and the storage server IP, and then the mount script, the user directory is mounted directly to the cloud desktop for users, until the user logs off cloud desktop automatically uninstalls storage.

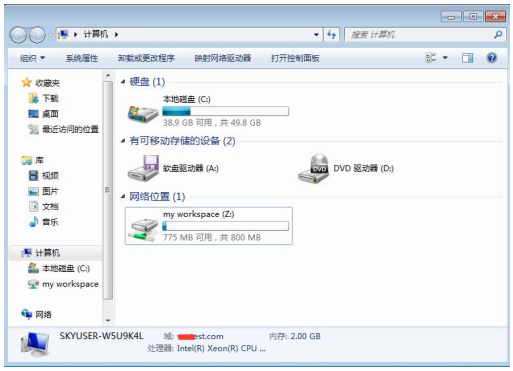


Figure 7: cloud storage and cloud desktop

III. EXPERIMENT AND DISCUSSION

Cloud desktop: Log virtual desktop, edit documents, browse pictures and other operations more smoothly

VNC console: Login virtual desktops need to wait around 1s, editing documents and browse graphics have delayed phenomenon

Scenario testing	Occupancy bandwidth	Average bandwidth	Delay
Cloud desktop	70-150Kbps	130Kbps	0.1s
VNC console	50-100Kbps	75Kbps	0.6s

IV. CONCLUSION

(1) User uses the cloud desktop, without entering any information; user can log in cloud desktop directly from the UI portal interface, achieving a landing automation.

(2) Cloud desktop use Microsoft's RDP transfer protocol, comparing to the CloudStack VNC console, faster, it is more stable and more efficient.

(3) When the user login cloud desktop, cloud disk was automatically mounted, the user can operate files on the cloud disk.

(4) Users can achieve the management of cloud desktop and cloud disk through the portal UI interface easily and quickly.

(5) Due to the high hardware limitations, the system has not been stress tested, currently only a few Dell servers, desktop deployment system for hundreds of it appeared to be inadequate.

(6) The system now supports only XenServer [10] and VSphere [11], and then we can consider adding other virtualization platforms, such as KVM [12], Hyper-V [13] and other platforms, making resource management more flexible and resilient, the real "cloud" Desktop Management platform.

(7) The system can only get CPU and Memory total cloud desktop cloud desktop performance for real-time monitoring (CPU and Memory usage) to be developed.

ACKNOWLEDGMENT

The authors acknowledge the financial supports by the National Key Technology Support Program (2012BAH17B03) and the Program Project of CUC (XNG1138, YXJS2012319, YXJS2012206, BY2012230, BE2013054 and JSWHCY-2013-(98)).

REFERENCES

[1] CloudStack. https://en.wikipedia.org/wiki/Apache_CloudStack

[2] VNC. https://en.wikipedia.org/wiki/Virtual_Network_Computing

[3] Active Directory. https://en.wikipedia.org/wiki/Active_Directory

[4] cloudstack.apache.org. http://docs.cloudstack.apache.org/projects/cloudstack-administration/en/4.5/.

[5] Cloud storage http://www.webopedia.com/TERM/C/cloud_storage.html.

[6] CloudStack 4.3 and LDAP Integration Setup. http://thehyperadvisor.com/2014/06/03/cloudstack-4-3-and-ldap-integration-setup/

[7] LDAP. http://www.gracion.com/server/whatldap.html

[8] MAC address. https://en.wikipedia.org/wiki/MAC_address

[9] Yang Li, Xiaoyan Yan, Quan Zhou, Qijuan GAO. Study on thin client and streaming video of the SBC mode for network teaching [A]. IN Proceedings of International Symposium on Information Technologies and Applications in Education[C].2008

[10] Citrix. Over view of XenServer[OL].http://www.citrix.com/products/xenserver/overview.html.2012.

[11] VSphere. https://en.wikipedia.org/wiki/VMware_vSphere.

[12] KVM. http://www.linux-kvm.org/page/Main_Page.

[13] Hyper-v. https://en.wikipedia.org/wiki/Hyper-V

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Fine-Grained and Dynamic MapReduce Task Scheduling Scheme for the Heterogeneous Cloud Environment

Yingchi Mao, Haishi Zhong, Longbao Wang
College of Computer and Information
Hohai University
Nanjing 211100, China
E-mail: maoyingchi@gmail.com

***Abstract*—MapReduce framework is becoming more and more popular in various applications. However, Hadoop is a seriously limited by its MapReduce scheduler which does not work well in the heterogeneous environment. LATE MapReduce scheduling algorithm takes heterogeneous environment into consideration. However, it falls short of solving the poor performance due to the static manner during computing the tasks progress. In order to improve the cluster performance in a heterogeneous cloud environment, FiGMR – a Fine-Grained and dynamic MapReeduce scheduling algorithm, is proposed. FiGMR can significantly reduce the tasks execution time and improve the resources utilization. FiGMR includes historical and real-time online information obtained from each node to select the appropriate parameters to find the real slow task dynamically. Meanwhile, in order to further improve the cluster performance, FiGMR classifies map nodes into high-performance map node and low-performance map node. FiGMR classifies slow tasks into slow map tasks and slow reduce tasks. Map/Reduce slow nodes means nodes which execute map/reduce tasks using a longer time than most other nodes. In this way, FiGMR launches backup map tasks on nodes which are high-performance map nodes.**

Keywords- Cloud computing; MapReduce scheduling; Hadoop; Fine-grained; Heterogeneous environment

I. INTRODUCTION

We are at the beginning of a Big Data era, how to efficiently process massive amounts of data has become an important issue. MapReduce is a distributed programming model for expressing distributed computation massive amounts of data and an execution framework for large-scale data processing on clusters. MapReduce framework is becoming more and more popular in various applications.

The initial MapReduce model was designed for off-line data processing [4]. However, it is now widely applied in heterogeneous, sharing and multi-user environments. At present, the MapReduce scheduling algorithms mainly have FIFO (First Input First Output), Fair Scheduler [1], Capacity Scheduler [2] and LATE [3] (Longest Approximate Time to End). The default MapReduce scheduler in Hadoop just considers scheduling in homogeneous environment, and cannot to find slow tasks which result in the execution time delay. LATE scheduling algorithm tried to find real slow tasks by computing remaining time of all the tasks. However, it does not find the real slow tasks and consider the different computing capacity in a cluster. Dynamic Proportional Scheduler [5] provides more job sharing and prioritization capability in scheduling and also results in increasing share of cluster resources and more differentiation in service levels of different jobs. Matei Zaharia et al. propose the delay scheduling algorithm [6] to address the conflict between data locality and fairness. However, the method takes fairness withered as the cost and it doesn't fit for the jobs which have large size or few slots per node. Zhang et al. propose the NKS algorithm [7] to improve the data locality of map tasks. However, it is based on the homogeneous environment. Heterogeneous clusters [8] consist of kinds of nodes with different performance characteristics in computing power, memory capacity and disk speed. So the homogeneous scheduling algorithms can't deal with deadline constraints efficiently in heterogeneous environment.

In order to improve the cluster performance in a heterogeneous cloud environment, FiGMR – a Fine-Grained and dynamic MapReeduce scheduling algorithm, is proposed. FiGMR can significantly reduce the tasks execution time and improve the resources utilization. FiGMR is inspired by facts that slow tasks prolong the execution time of the whole job and nodes requires various time to complete the same tasks due to their heterogeneousness. For example, there are different capacities of computation, disk I/O, memory, and communication in a cluster. Although Hadoop and LATE can launch backup tasks for the slow tasks, they cannot determine the appropriate tasks which are really prolong the execution time of the whole job. The reason is that they always adopts a static way to find the slow tasks without considering the dynamic capacities among the different nodes during the tasks execution. On the contrary, FiGMR includes historical and real-time online information obtained from each node to select the appropriate parameters to find the real slow task dynamically. Meanwhile, in order to further improve the cluster performance, FiGMR classifies map nodes into high-performance map node and low-performance map node. FiGMR classifies slow tasks into slow map tasks and slow reduce tasks. Map/Reduce slow nodes means nodes which execute map/reduce tasks using a longer time than most other nodes. In this way, FiGMR launches backup map tasks on nodes which are high-performance map nodes.

The contribution of this paper are as follows:

1) According to historical tasks of all the finished jobs and online tasks of the current job, FiGMR establishes a fine-grained and dynamic node performance capacities model to evaluate reference performance and real-time performance for different types of tasks.

2) To improve task data locality, FiGMR adopts a data distribution strategy based on nodes' performance in the heterogeneous environment.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.46

3) To accurately estimate the tasks completion time and to determine the backward tasks, FiGMR classifies map nodes into high-performance map node and low-performance map node. FiGMR classifies slow tasks into slow map tasks and slow reduce tasks. FiGMR launches backup map tasks on nodes which are high-performance map nodes.

The rest of this paper is organized as follows. Section II presents the performance computation model. In section III, a fine-grained and dynamic MapReduce scheduling algorithm for the heterogeneous environments is proposed. Section IV presents the experiment results of data locality, task completion time and task scheduling. We conclude the paper and future work in Section V.

II. NODE PERFORMANCE COMPUTATION MODEL

In a MapReudce job, the input data to Map task is the data fragment stored in the distributed file system, and the input data to Reduce task is the intermediate results from the Map tasks. Based on the historical and online information of tasks and nodes, the system can compute the node capacities from the different following definitions.

Definition 1: Map task execution rate $MapRate$ represents the size of processed data in a unit time. That is to say, $MapRate$ means the ratio of the processed data size to the processing time.

$$MapRate = \frac{Size\_data}{Time\_M} \tag{1}$$

Definition 2: For one node N_i and one job Job_j, the execution speed of map task of Job_j on the node N_i is calculated as the average execution speed of s map tasks of Job_j on the node N_i.

$$MapRate\_NodeJob_{ij} = \frac{\sum_{k=1}^{s} MapRate_k}{s} \tag{2}$$

Definition 3: High-performance map node. If the execution speed of map tasks of Job_j on the node N_i is greater than the average execution speed of map tasks on all of the nodes in a cluster, the node N_i can be called as the high-performance map node. The default value of $Threshold\_slownode$ is 25% in LATE algorithm.

$$MapRate\_NodeJob_{ij} \geq (1 - Threshold\_slownode) \times \frac{\sum_{k=1}^{M} MapRate\_NodeJob_{kj}}{M} \tag{3}$$

According to the above definition, $MapRate\_NodeJob_{ij}$ can be used to as the real-time map performance.

Definition 4: For any one node N_i, the execution speed of the historical map tasks on N_i, $MapRate\_Node_i$, can be calculated as the average execution speed of s completed map tasks on the node N_i.

$$MapRate\_Node_i = \frac{\sum_{k=1}^{s} MapRate\_NodeJob_{ik}}{s} \tag{4}$$

Definition 5: Reference high-performance map node. If the execution speed of historical map tasks on the node N_i is greater than the threshold $Threshold\_slownode$, average execution speed of historical map tasks on all of the nodes in a cluster, the node N_i can be called as the reference high-performance map node.

$$MapRate\_Node_i \geq (1 - Threshold\_slownode) \times \frac{\sum_{k=1}^{M} MapRate\_Node_k}{M} \tag{5}$$

Definition 6: For any one node N_i, the reference map capacity of node N_i, $MapCapacity_i$, can be represented as the rate of the execution speed of the historical map tasks on N_i, $MapRate\_Node_i$, to the sum of the execution speed of the historical map tasks on all of nodes in a cluster.

$$MapCapacity_i = \frac{MapRate\_Node_i}{\sum_{k=1}^{M} MapRate\_Node_k} \tag{6}$$

In a similar way, we can obtain the corresponding definitions for the reduce tasks and reduce nodes.

III. FIGMR: FINE-GRAINED AND DYNAMIC MAPREDUCE SCHEDULING

A. Main Idea of FiGMR Algorithm

FiGMR is developed based on LATE MapReduce scheduling algorithm. However, FiGMR can obtain more accurate the progress score of all the tasks by using information and real-time online capacities of each node. By using accurate the progress score, FiGMR can find the real slow map and reduce tasks and launch the corresponding backup tasks in order to decrease the execution time compared with Hadoop and LATE.

FiGMR scheduler tasks three steps to realize the fine-grained and dynamic MapReduce task scheduling. First, new tasks are submitted to a stack of MapReduce tasks. All of the task trackers obtain the new tasks from the stack based on the data locality. Second, the task trackers collect comprehensive node's performance, and compute the reference performance and online performance for all the tasks running on the corresponding nodes. Next, FiGMR should determine which tasks are slow map tasks, slow reduce tasks, high-performance map node and high-performance reduce node. Consequently, these slow tasks are inserted into corresponding queue of slow tasks. Meanwhile, if the stack of new tasks is empty, the task tracker will start to lookup one of slow tasks in the queue, and launch backup task. Only when the task tracker is not a map/reduce slow node, it can launch backup tasks for the map tasks or reduce tasks

B. Computing node's capacity based on the historical and real-time information

As a data-intensive computing framework, most of MapReduce's jobs are toward the massive data processing. The task's response time and completion time is related to the node's CPU, disk I/O, memory and etc.

Two purposes of node performance computation: (1) To optimize the data distribution in order to improve the data locality; (2) To improve the evaluation accuracy of task remaining time in the heterogeneous environments.

In this paper, we use the comprehensive criterion to represent the data processing performance for each

physical machine node. The criterion of node capacity includes all memory intensity, I/O intensity, and CPU intensity jobs. In the heterogeneous Cloud environment, due to nodes with different processing capacity, the processing capacity is introduced.

First, *tasks trackers* collect historical information from the nodes where they are running on. The historical information includes historical values of *M1, M2, R1, R2* and *R3*. Then, *tasks trackers* select the appropriate value of *M1, M2, R1, R2* and *R3* according to historical and online information. Consequently, *tasks trackers* collect values of *M1, M2, R1, R2* and *R3* according to the real running information after the tasks finished. Finally, *tasks trackers* update these historical information stored on the nodes.

C. Finding Slow Tasks

In MapReduce, the job's execution progress includes Map and Reduce stage. The Job's completion time contains Map execution time and Reduce execution time.

The map task execution can be further split into two stage. One is to read input data and execute map function, called as "map function execution" stage. The other is to resort the intermediate results and merge the results, called as "resort" stage. The weights of two Map stages are $M1$ and $M2$, respectively. The reduce task can be divided into "copy data", "sort", and "merge results". The weight of three Reduce stages are $R1$, $R2$ and $R3$ respectively. Therefore, $M1+M2=1$ and $R1+R2+R3=1$. Adopting the node performance computing model, it can compute the weights.

Suppose M_{finish} is the number of input data tuples which has been processed in a task, and M_{all} is the number of overall data tuples in the task. For the current stage of processing S ($S=0,1,2$), the process score is $StageScore$. The $StageScore$ in a certain stage can be computed with $StageScore=\frac{M_{finish}}{M_{all}}$.

The progress score of map tasks and reduce tasks, $MProgressScore$ and $RProgressScore$, can be computed according the Eq. 3 And 4.

$$MProgressScore=\begin{cases} M1\times StageScore & S=0 \\ M1+M2\times StageScore & S=1 \end{cases} \quad (7)$$

$$RProgressScore=\begin{cases} R1\times StageScore & S=0 \\ R1+R2\times StageScore & S=1 \\ R1+R2+R3\times StageScore & S=2 \end{cases} \quad (8)$$

FiGMR computes *PS* more accurate than Hadoop and LATE, Because *M1, M2, R1, R2* and *R3* are selected according to historical information. However, in Hadoop and LATE, *M1, M2, R1, R2* and *R3* are 1, 0, 0.33, 0.33, 0.34 respectively, which cannot adaptive to different environment. After getting exact *ProgressScore*, *FiGMR* computes the remaining time of all the running tasks, *TTE*, according to the Eq. 9 and Eq. 10. By this way, *FiGMR* finds real slow tasks and launches backup tasks for these slow tasks on fast nodes of this kind of tasks consequently.

$$ProgressRate = ProgressScore / Tracker \quad (9)$$

$$TTE = (1.0 - ProgressScore) / ProgressRate \quad (10)$$

D. Launching backup tasks

If there are slow tasks, and Eq. 14 is fulfilled, a backup task can be launched when some of *tasks trackers* are free. The probability of backup tasks is used to define the maximum proportion of backup tasks in all the tasks. Suppose the number of backup tasks is *BackupNum*, the number of all the running tasks is *TaskNum*. The Eq. 11 must be fulfilled in the system.

$$BackupNum < BP * TaskNum \quad (11)$$

FiGMR can launch a backup task for slow task Ti and the number of backup tasks is less than the maximum number of backup tasks according to the Eq. 11.

IV. EVALUATION

A. Experiments Setup

We establish experiments by a Hadoop cluster with seven nodes with different computing capacity. All of the cluster configures are listed in Table II. The LATE and FiGMR are implemented based on Hadoop 1.0.0.

TABLE I. EXERIMENTAL SETUP

Nodes	CPU	RAM	Storage
Master	32 Core 3.6GHz	96 GB	2 TB
Slave1	1 Core 2.3GHz	1 GB	250 GB
Slave2	2 Core 2.6GHz	4 GB	500 GB
Slave3	4 Core 3.2GHz	16 GB	750 GB
Slave4	1 Core 2.3 GHz	2 GB	250 GB
Slave5	2 Core 2.9 GHZ	4 GB	500 GB
Slave6	2 Core 3.2 GHz	8 GB	500 GB

B. Data locality

Data locality is a key performance of Hadoop MapReduce jobs. In the experiments, we run Sort and WordCount program to measure the proportion of the data locality and the corresponding tasks completion time. The amount of input data tuples varies from 0.3GB to 5.0GB. Since LATE scheduling algorithm adopts the same data distribution with Hadoop, we compare the job's data locality performance with Hadoop default scheduler and FiGMR scheduler.

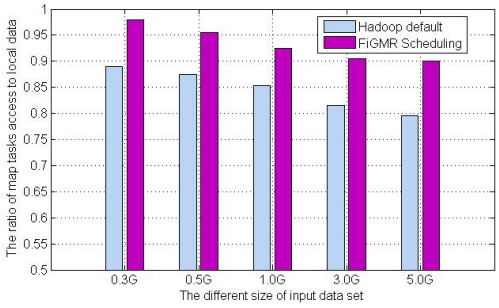


Figure 1. The proportion of data locality

The results of the proportion of the data locality and the corresponding tasks completion time are shown in Fig. 1 and 2, respectively. When the number of data block is

smaller than the number of slots for map tasks in the cluster, it adopts the data distribution with the priority of high-performance map nodes. Thus, almost map tasks can be run on the high-performance map nodes and access the data locality. As shown in Fig. 1 and 2, the improved data distribution strategy can improve the proportion of data locality 10.1% and 9.1%, and reduce the job's completion time 15.7% and 14.9%, while comparing with Hadoop default scheduling on average. On the other hand, when the number of data block is greater than the number of slots in the cluster, FiGMR algorithm can improve the proportion of data locality 9.4%, 11.0%, and 12.5% while comparing the Hadoop default scheduling algorithm, in running job of Sort. The reason is that FiGMR scheduling algorithm can effectively reduce the network overhead when the data is forwarded from the low-performance map nodes to the high-performance map nodes. It can make full use of the computing capacity on the high-performance nodes. FiGMAR can reduce the job's completion time 14.5%, 14.8%, and 17.1%, respectively.

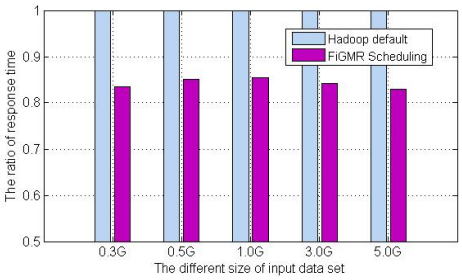


Figure 2. The completion time of tasks

C. *Tasks Scheduling Performance*

In the FiGMR scheduling algorithm, to improve the accuracy of slow tasks detection, it adopts different weights for those five steps to compute the progress scores of the map/reduce task. In the experiments, we evaluate the tasks completion time of Hadoop, LATE and FiGMR algorithm in the average, best and worst case, when the input data tuples is 3.0G.

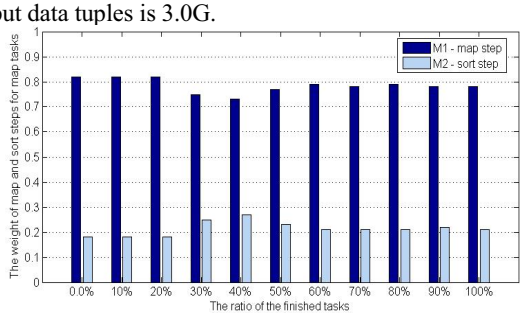


Figure 3. The weight of Map tasks

Fig. 3 and 4 shows the different weights in the different completion ratio of tasks. As shown in Fig. 3 and 4, when the task completion ratio in a node is from 0%-20%, we use the historical information to calculate weights as the reference weights. When the completion ratio varies from 30% -80%, the online information is used to calculate the weights for the real-time tasks. When the completion ratio is greater than 90%, the value of different weights is convergent.

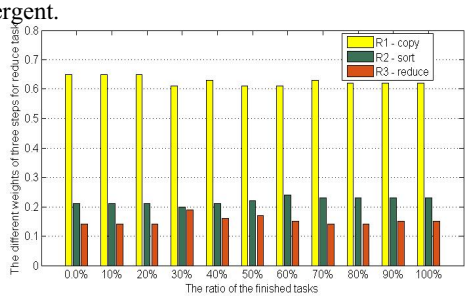


Figure 4. The weight of Reduce tasks

V. CONCLUSION AND FUTURE WORK

In order to improve the cluster performance in a heterogeneous cloud environment, FiGMR – a Fine-Grained and dynamic MapReeduce scheduling algorithm, is proposed. FiGMR can significantly reduce the tasks execution time and improve the resources utilization. FiGMR includes historical and real-time online information obtained from each node to select the appropriate parameters to find the real slow task dynamically.

ACKNOWLEDGMENT

This research is partially supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under Grant No. 2013BAB06B04; Key Technology Project of China Huaneng Group under Grant No.HNKJ13-H17-04; Science and Technology Program of Yunnan Province under Grant No. 2014GA007; Nature Science Fund of Jiangsu Province under Grant No. BK20130852.

REFERENCES

[1] Zaharia, M., Konwinski, A., Joseph, A.D., Katz, R., Stoica, I.: Improving MapReduce performance in heterogeneous environments. In: Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation, pp. 29–42 (2008)

[2] Zaharia, M., Borthakur, D., Sarma, J.S., Elmeleegy, K., Shenker, S., Stoica, I.: Job scheduling for multi-user MapReduce clusters. (2009). http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-55.pdf. Accessed 1 March 2012

[3] http://hadoop.apache.org/mapreduce/docs/r0.21.0/capacity_scheduler.html (2011). Accessed 1 March 2012

[4] The Apache Software Foundation: Hadoop (2012). http://hadoop.apache.org. Accessed 1 March 2012

[5] Sandholm, T., Lai, K.: Dynamic proportional share scheduling in hadoop. In: Proceedings of the 15th Workshop on Job Scheduling Strategies for Parallel Processing, pp. 110–131 (2010)

[6] Zaharia, M., Borthakur, D., Sarma, J.S., Elmeleegy, K., Shenker, S., Stoica, I.: Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In: Proceedings of the 5th European Conference on Computer Systems, pp. 265–278 (2010)

[7] Zhang, X.H., Zhong, Z.Y., Feng, S.Z., Tu, B.B., Fan, J.P.: Improving data locality of MapReduce by scheduling in homogeneous computing environments. In: IEEE 9th International Symposium on Parallel and Distributed Processing with Applications, pp. 120–126 (2011). doi:10.1109/ISPA.2011.14

[8] Qin, X., Jiang, H., Manzanares, A., Ruan, X., Yin, S.: Dynamic load balancing for IO-intensive applications on clusters. ACM Trans. Storage 5(3), 1–38 (2009). doi:10.1145/1629075.1629078

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Cryptanalysis of Two Tripartite Authenticated Key Agreement Protocols

Yang Lu
College of Computer and Information Engineering,
Hohai University
Nanjing, Jiangsu 211100 - China
luyangnsd@163.com

Quanling Zhang, Jiguo Li
College of Computer and Information Engineering,
Hohai University
Nanjing, Jiangsu 211100 - China
zhangquanling99@163.com, ljg1688@163.com

***Abstract*—The tripartite authenticated key agreement (AKA) protocol is crucial in providing data confidentiality to subsequent communications among three parties over an insecure public network. In recent years, several identity-based and certificateless tripartite AKA protocols have been proposed. Unfortunately, most of them are insecure. In this paper, we analyze an identity-based tripartite AKA protocol and a certificateless tripartite AKA protocol. Our cryptanalysis shows that one neither meets the desirable security properties.**

Keywords-authenticated key agreement; identity-based; certificateless cryptosystme; tripartite; bilinear pairing

I. INTRODUCTION

Key agreement protocol is an important cryptographic primitive which enables two or more communicating entities to establish a shared session key over an open network. The session key can be used for encryption purpose in order to provide confidentiality. This kind of key agreement protocols are only secure against passive adversaries. In the real world, the adversary may mount more powerful attacks such as by impersonating one party to communicate with another party. Thus, the notion of authenticated key agreement (AKA) has been proposed to defeat such active adversaries. A tripartite AKA protocol allows three parties to securely establish a shared secret key for their communications in the present of an adversary. Thus, secure tripartite AKA protocols serve as basic building block for constructing secure, complex, higher-level protocols.

In 2013, Xiong *et al.* [1] proposed a new identity-based tripartite AKA protocol with provable security. Unfortunately, Lin *et al.* [2] found that Xiong *et al.*'s protocol is insecure against insider replay attack. In this paper, we propose a different attack which is named key compromise impersonation (KCI) attack. In 2012, Xiong *et al.* [3] proposed a certificateless tripartite AKA protocol. However, Sun *et al.* [4] mounted a KCI attack on Xiong *et al.*'s protocol [3] and proposed an improved protocol to make up the security weakness. Unfortunately, we find that Sun *et al.*'s improved protocol [4] is also insecure against insider replay attack.

The rest of the paper is organized as follows. Section II introduces some preliminaries. We review Xiong *et al.*'s protocol [1] and present the KCI attack in Section III. In Section IV, we review Sun *et al.*'s improved protocol [4] and show that it is vulnerable to the insider replay attack. Finally, a conclusion is given in Section V.

II. BILINEAR PAIRING AND SECURITY PROPERTIES

A. Bilinear pairing

Let q be a prime number, G_1 an additive cyclic group of prime order q and G_2 a multiplicative cyclic group of the same order. A mapping $e: G_1 \times G_1 \to G_2$ is called a bilinear map if it satisfies the following properties:

(1)Bilinearity: $e(aU, bV) = e(U,V)^{ab}$ for all $U, V \in G_1$ and $a, b \in Z_q^*$.

(2) Non-degeneracy: There exists $U, V \in G_1$ such that $e(U,V) \neq 1$.

(3) Computability: $e(U,V)$ can be efficiently computed for all $U, V \in G_1$.

B. Security properties

It is desirable for AKA protocols to possess the security properties as claimed [5, 6, 7]. Here, we highlight the security properties of tripartite AKA protocols as follows.

(1) *Known-key security*: Even if some of the session keys of a given protocol are leaked, an adversary should be unable to learn other session keys.

(2) *Forward secrecy*: If the private keys of one or more participants are compromised, the secrecy of previously established session keys should not be affected. This security property includes three types: (*a*) *Perfect forward secrecy*: even if all participants' private keys are compromised, *forward secrecy* must be preserved; (*b*) *Partial forward secrecy*: even if one or more participants but not all participants' private keys are compromised, *forward secrecy* must be preserved; (*c*) *PKG*(*KGC*) *forward secrecy*: even if the PKG's or KGC's master key is compromised, *forward secrecy* must be preserved.

(3) *Unknown key-share resilience*: An adversary should be unable to force a group of participants to share a key with him, whereas in reality they are sharing a key with another participant.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.47

(4) *Basic impersonation attacks resilience*: An adversary should be unable to impersonate a participant if it does not know the participant's private key.

(5) *Key compromise impersonation resilience*: If the private key of a participant A is compromised, an adversary can be able to impersonate A to other participants but can not impersonate others to the participant A.

(6) *Key control*: The session key should be determined jointly by the three participants. An adversary should be unable to force the participants to accept a pre-selected value as the current session key.

Note that a tripartite AKA protocol should also be secure against message replay attack and reflect attack, etc. In this paper, we show that the two tripartite AKA protocols do not satisfy the desirable security properties.

III. CRYPTANALYSIS OF XIONG ET AL.'S PROTOCOL

In this section, we review Xiong *et al.*'s protocol [1] as follows. Then, we give an attack which is different from shown in the literature [2].

A. Xiong et al.'s protocol

(1) ***Setup*:** Given a security parameter $k \in Z$, the algorithm works as follows:

1. Run the parameter generator on input k to generate a prime q, two groups G_1, G_2 of prime order q, a generator P of G_1 and an admissible mapping $\hat{e}: G_1 \times G_1 \to G_2$.
2. Select a master-key $x \in_R Z_q^*$, and compute $P_{pub} = xP$.
3. Choose cryptographic hash functions H_1: $\{0,1\}^* \times G_1 \to Z_q^*$ and H_2: $\{0,1\}^* \times \{0,1\}^* \times \{0,1\}^* \times G_1^8 \to \{0,1\}^k$.

Finally the PKG's master-key x is kept secret and the system parameters $\{k, q, \hat{e}, G_1, G_2, P, P_{pub}, H_1, H_2\}$ are published.

(2) ***Private Key Extraction*:** Given a user's identity $ID_U \in \{0,1\}^*$, PKG first chooses at random $r_U \in Z_q^*$, computes $R_U = r_U P$, $h = H_1(ID_U \| R_U)$ and $s_U = (r_U + hx)^{-1}$. It then sets this user's private key (s_U, R_U), and transmits it to user ID_U secretly.

It is easy to see that user ID_U can validate his long-term key by checking whether the equation $s_U(R_U + H_1(ID_U \| R_U)P_{pub}) = P$ holds. The long-term key is valid if the equation holds and vice versa.

(3) ***Key Agreement:*** The message flows and computations of a protocol run are described below.

1. A, B, C: choose $a, b, c \in Z_q^*$
2. $A \to B, C: \{ID_A, R_A\}$

$B \to A: \{ ID_B, R_B, T_{BA} = b(R_A + H_1(ID_A \| R_A)P_{pub}) \}$

$C \to A: \{ ID_C, R_C, T_{CA} = c(R_A + H_1(ID_A \| R_A)P_{pub}) \}$

$A \to B: T_{AB} = a(R_B + H_1(ID_B \| R_B)P_{pub})$

$A \to C: T_{AC} = a(R_C + H_1(ID_C \| R_C)P_{pub})$

$B \to C: \{ID_B, R_B\}$

$C \to B: \{ ID_C, R_C, T_{CB} = c(R_B + H_1(ID_B \| R_B)P_{pub}) \}$

$B \to C: T_{BC} = b(R_C + H_1(ID_C \| R_C)P_{pub})$

3. A computes:

$K_{ABC}^1 = aP + s_A T_{BA} + s_A T_{CA} = aP + bP + cP = (a+b+c)P$

$K_{ABC}^2 = \hat{e}(s_A T_{BA}, s_A T_{CA})^a = \hat{e}(bP, cP)^a = \hat{e}(P,P)^{abc}$

B computes:

$K_{ABC}^1 = bP + s_B T_{AB} + s_B T_{CB} = bP + aP + cP = (a+b+c)P$

$K_{ABC}^2 = \hat{e}(s_B T_{AB}, s_B T_{CB})^b = \hat{e}(aP, cP)^b = \hat{e}(P,P)^{abc}$

C computes:

$K_{ABC}^1 = cP + s_C T_{AC} + s_C T_{BC} = cP + aP + bP = (a+b+c)P$

$K_{ABC}^2 = \hat{e}(s_C T_{AC}, s_C T_{BC})^c = \hat{e}(aP, bP)^c = \hat{e}(P,P)^{abc}$

After the protocol has finished, all three entities share the session key which is computed as

$K = H_2(ID_A \| ID_B \| ID_C \| T_{AB} \| T_{AC} \| T_{BA} \| T_{BC} \| T_{CA} \| T_{CB} \| K_{ABC}^1 \| K_{ABC}^2)$

B. Key compromise impersonation attack on Xiong et al.'s protocol

Suppose that A's private key (s_A, R_A) and B's private key (s_B, R_B) have been compromised. Obviously, the adversary E is able to impersonate the corrupted party to any other party. With A's private key and B's private key at hand, the adversary E attempts to establish a valid session key with A and B by masquerading as another legitimate entity C. Note that, with A's private key, E can initiate a protocol run with B and C by impersonating A to obtain R_C.

A concrete KCI attack by the adversary E against Xiong *et al.*'s protocol is described below. Where, $E(C)$ denotes that E is impersonating C.

1. $A, B, E(C)$: choose $a, b, c' \in Z_q^*$
2. $A \to B, E(C): \{ID_A, R_A\}$

$B \to A: \{ ID_B, R_B, T_{BA} = b(R_A + H_1(ID_A \| R_A)P_{pub}) \}$

$E(C) \to A: \{ ID_C, R_C, T_{E(C)A} = c'(R_A + H_1(ID_A \| R_A)P_{pub}) \}$

$A \to E(C): T_{AE(C)} = a(R_C + H_1(ID_C \| R_C)P_{pub})$

$B \to E(C): \{ID_B, R_B\}$

$E(C) \to B: \{ ID_C, R_C, T_{E(C)B} = c'(R_B + H_1(ID_B \| R_B)P_{pub}) \}$

$B \to E(C): T_{BE(C)} = b(R_C + H_1(ID_C \| R_C)P_{pub})$

3. A and B compute the session key according to the protocol specification. $E(C)$ computes the session key as follows.

$K_{ABC}^1 = c'P + s_B T_{AB} + s_A T_{BA} = c'P + aP + bP$
$= (a+b+c')P$

$$K^2_{ABC} = \hat{e}(s_B T_{AB}, s_A T_{BA})^{c'} = \hat{e}(aP, bP)^{c'} = \hat{e}(P,P)^{abc'}$$
$$K = H_2(ID_A \| ID_B \| ID_C \| T_{AB} \| T_{AE(C)} \| T_{BA} \| T_{BE(C)} \| T_{E(C)A} \| T_{E(C)B} \| (a+b+c')P \| \hat{e}(P,P)^{abc'})$$

Hence, the adversary E successfully establishes a session key K with entity A and B while A and B believe they are sharing the key with entity C. Thus, Xiong *et al.*'s protocol [1] is vulnerable to a KCI attack.

IV. CRYPTANALYSIS OF SUN ET AL.'S IMPROVED PROTOCOL

In this section, we review Sun *et al.*'s improved protocol [4] as follows. Then, we give an insider replay attack.

A. Sun et al.'s protocol

(1) ***Setup*:** Given a security parameter $k \in Z$, the algorithm works as follows.

1. It runs the parameter generator on input k to generate a prime q, two groups G_1, G_2 of prime order$q$, a generator P of G_1 and an admissible mapping $\hat{e}: G_1 \times G_1 \to G_2$.
2. It selects a master-key $x \in_R Z_q^*$, and computes $P_{pub} = xP$.
3. It chooses cryptographic hash functions H_1: $\{0,1\}^* \times G_1 \to Z_q^*$ and H_2: $\{0,1\}^* \times \{0,1\}^* \times \{0,1\}^* \times G_1^{10} \times G_2^2 \to \{0,1\}^k$. Finally the KGC's master-key x is kept secret and the system parameters $\{k, q, \hat{e}, G_1, G_2, P, P_{pub}, H_1, H_2\}$ are published.

(2) ***PartialKeyGen*:** Given a user's identity $ID_U \in \{0,1\}^*$, the KGC first chooses at random $r_U \in Z_q^*$, and computes $R_U = r_U P$, $h = H_1(ID_U \| R_U)$ and $s_U = (r_U + hx)^{-1}$. It then sets this user's partial private key(s_U, R_U) and transmits it to user ID_U secretly.

It is easy to see that user ID_U can validate his partical private key by checking whether the equation $s_U(R_U + H_1(ID_U \| R_U)P_{pub}) = P$ holds. The partial key is valid if the equation holds, and vice versa.

(3) ***UserKeyGen*:** User ID_U selects value $x_U \in_R Z_q^*$ as his secret key usk_U, and computes $upk_U = x_U P$ as his public key .

(4) ***Key Agreement*:** Assume that an entity A with identity ID_A has full private key(s_A, R_A, x_A) and public key upk_A, an entity B with identity ID_B has full private key (s_B, R_B, x_B) and public key upk_B, and an entity C with identity ID_C has full private key(s_C, R_C, x_C) and public key upk_C. The message flows and computaions of a protocol run are described below.

1. A, B, C: choose a, b, $c \in_R Z_q^*$
2. $A \to B, C$: $\{ID_A, upk_A, R_A\}$

   $B \to A$: { ID_B , upk_B , R_B , $T_{BA} = b(R_A + H_1(ID_A \| R_A)P_{pub})$ }

   $C \to A$: { ID_C , upk_C , R_C , $T_{CA} = c(R_A + H_1(ID_A \| R_A)P_{pub})$ }

   $A \to B$: $T_{AB} = a(R_B + H_1(ID_B \| R_B)P_{pub})$

   $A \to C$: $T_{AC} = a(R_C + H_1(ID_C \| R_C)P_{pub})$

   $B \to C$: $\{ID_B, upk_B, R_B\}$

   $C \to B$: { ID_C , upk_C , R_C , $T_{CB} = c(R_B + H_1(ID_B \| R_B)P_{pub})$ }

   $B \to C$: $T_{BC} = b(R_C + H_1(ID_C \| R_C)P_{pub})$
3. A computes:

$$K^1_{ABC} = aP + s_A T_{BA} + s_A T_{CA} = (a+b+c)P$$
$$K^2_{ABC} = \hat{e}(s_A T_{BA} + R_B + H_1(ID_B \| R_B)P_{pub}, s_A T_{CA} + R_C + H_1(ID_C \| R_C)P_{pub})^{a+s_A^{-1}} = \hat{e}(P,P)^{(a+s_A^{-1})(b+s_B^{-1})(c+s_C^{-1})}$$
$$K^3_{ABC} = \hat{e}(s_A T_{BA} + upk_B, s_A T_{CA} + upk_C)^{a+x_A} = \hat{e}(P,P)^{(a+x_A)(b+x_B)(c+x_C)}$$

   B computes:

$$K^1_{ABC} = bP + s_B T_{AB} + s_B T_{CB} = (a+b+c)P$$
$$K^2_{ABC} = \hat{e}(s_B T_{AB} + R_A + H_1(ID_A \| R_A)P_{pub}, s_B T_{CB} + R_C + H_1(ID_C \| R_C)P_{pub})^{b+s_B^{-1}} = \hat{e}(P,P)^{(a+s_A^{-1})(b+s_B^{-1})(c+s_C^{-1})}$$
$$K^3_{ABC} = \hat{e}(s_B T_{AB} + upk_A, s_B T_{CB} + upk_C)^{b+x_B} = \hat{e}(P,P)^{(a+x_A)(b+x_B)(c+x_C)}$$

   C computes:

$$K^1_{ABC} = cP + s_C T_{AC} + s_C T_{BC} = (a+b+c)P$$
$$K^2_{ABC} = \hat{e}(s_C T_{AC} + R_A + H_1(ID_A \| R_A)P_{pub}, s_C T_{BC} + R_B + H_1(ID_B \| R_B)P_{pub})^{c+s_C^{-1}} = \hat{e}(P,P)^{(a+s_A^{-1})(b+s_B^{-1})(c+s_C^{-1})}$$
$$K^3_{ABC} = \hat{e}(s_C T_{AC} + upk_A, s_C T_{BC} + upk_B)^{c+x_C} = \hat{e}(P,P)^{(a+x_A)(b+x_B)(c+x_C)}$$

After the protocol has finished, all three entities share the session key, which is computed as $K = H_2(ID_A \| ID_B \| ID_C \| upk_A \| upk_B \| upk_C \| T_{AB} \| T_{AC} \| T_{BA} \| T_{BC} \| T_{CA} \| T_{CB} \| K^1_{ABC} \| K^2_{ABC} \| K^3_{ABC})$.

B. Insider replay attack on Sun et al.'s protocol

Suppose three entities A, B and C have completed one protocol run *Round*1. It is reasonable to assume that A can obtain R_C when C participants in the protocol run *Round*1. Then, A can fool B into believing that they have participated in a protocol run with C, but in fact C has not been active. The attack is described as follows.

The insider entity A firstly initiates a new protocol run *Round*2 with the entity B and at the same time impersonates

the entity C. Then, A sends his message to B and replays C's message produced in the previous protocol run *Round*1. Let $A(C)$ denote the entity C impersonated by A. They perform as follows.

1. A: choose $a, c' \in_R Z_q^*$, B: choose $b \in_R Z_q^*$
2. $A \rightarrow B, A(C)$: $\{ID_A, upk_A, R_A\}$

   $B \rightarrow A$: $\{ ID_B , upk_B , R_B , T_{BA} = b(R_A + H_1(ID_A \| R_A)P_{pub}) \}$

   $A(C)$ broadcasts: ID_C , upk_C , R_C , $T_{A(C)A} = c'(R_A + H_1(ID_A \| R_A)P_{pub})$

   $A \rightarrow B$: $T_{AB} = a(R_B + H_1(ID_B \| R_B)P_{pub})$

   A broadcasts: $T_{AA(C)} = a(R_C + H_1(ID_C \| R_C)P_{pub})$

   $B \rightarrow A(C)$: $\{ID_B, upk_B, R_B\}$

   $A(C) \rightarrow B$: $\{ ID_C , upk_C , R_C , T_{A(C)B} = c'(R_B + H_1(ID_B \| R_B)P_{pub}) \}$

   $B \rightarrow A(C)$: $T_{BA(C)} = b(R_C + H_1(ID_C \| R_C)P_{pub})$
3. B computes:

$$K_{ABC}^1 = bP + s_B T_{AB} + s_B T_{A(C)B} = (a + b + c')P$$

$$K_{ABC}^2 = \hat{e}(s_B T_{AB} + R_A + H_1(ID_A \| R_A)P_{pub}, s_B T_{A(C)B} + R_C + H_1(ID_C \| R_C)P_{pub})^{b+s_B^{-1}} = \hat{e}(P,P)^{(a+s_A^{-1})(b+s_B^{-1})(c'+s_C^{-1})}$$

$$K_{ABC}^3 = \hat{e}(s_B T_{AB} + upk_A, s_B T_{A(C)B} + upk_C)^{b+x_B} = \hat{e}(P,P)^{(a+x_A)(b+x_B)(c'+x_C)}$$

   A computes:

$$K_{ABC}^1 = aP + s_A T_{BA} + s_A T_{A(C)A} = (a + b + c')P$$

$$K_{ABC}^2 = \hat{e}(s_A T_{BA} + R_B + H_1(ID_B \| R_B)P_{pub}, s_A T_{A(C)A} + R_C + H_1(ID_C \| R_C)P_{pub})^{a+s_A^{-1}} = \hat{e}(P,P)^{(a+s_A^{-1})(b+s_B^{-1})(c'+s_C^{-1})}$$

$$K_{ABC}^3 = \hat{e}(s_A T_{BA} + upk_B, s_A T_{A(C)A} + upk_C)^{a+x_A} = \hat{e}(P,P)^{(a+x_A)(b+x_B)(c'+x_C)}$$

Thus, A successfully fool B into believing that they have participated in a protocol run with C, but in fact C has not been active.

V. CONCLUSION

In this paper, we have analyzed the security of Xiong *et al.*'s identity-based tripartite AKA protocol [1] and Sun *et al.*'s certificateless tripartite AKA protocol [4]. We show that these protocols do not possess the desirable security properties. How to design a secure identity-based or certificateless tripartite AKA protocol to prevent all kinds of attacks is still interesting and challenging.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments and suggestions. The research was supported by the NSF-China-61272542.

REFERENCES

[1] H. Xiong, Z. Chen, and F. Li, "New identity-based Three-Party Authenticated Key Agreement Protocol with Provable Security", Journal of Network and Computer Applications, vol.36, Mar. 2013, pp. 927-932, doi:10.1016/j.jnca.2012.10.001.

[2] X. Lin and L. Sun, "Weakness of Several identity-based Tripartite Authenticated Key Agreement Protocols", (Cryptology Eprint Archive: Report 2013/848), https://eprint.iacr.org/2013/848.pdf, unpublished.

[3] H. Xiong, Z. Chen, and F. Li, "Provably Secure and Efficient Certificateless Authenticated Tripartite Key Agreement Protocol", Mathematical and Computer Modeling, vol.55, Feb. 2012, pp. 1213-1221, doi:10.1016/j.mcm.2011.10.001.

[4] H. Sun, Q. Wen, H. Zhang, Z. Jin and W. Li, "Cryptanalysis and Improvement of Two Certificateless Three-Party Authenticated Key Agreement Protocols", 2013, (arXiv:1301.5091), http://arxiv.org/ pdf/ 1301.5091v1.pdf, unpublished.

[5] M. Hölbl, T. Welzer and, B. Brumen," Two Proposed identity-based Three Party Authenticated Key Agreement Protocols from Pairings", Computers & Security, vol. 29, Mar. 2010, pp. 244–252, doi:10.1016 /j.cose.2009.08.006.

[6] P. Nose, "Security Weaknesses of Authenticated Key Agreement Protocols", Information Processing Letters, vol.111, Jul. 2011, pp. 687–696, doi:10.1016/j.ipl.2011.04.007.

[7] Z. Tan, "An Enhanced Three-Party Authentication Key Exchange Protocol for Mobile Commerce Environments", Journal of Communications, vol.5, May. 2010, pp. 436–443, doi:10. 4304 /jcm. 5.5.436-443.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Schnorr ring signature scheme with designated verifiability

Xin Lv, Feng Xu, Ping Ping, Xuan Liu
College of Computer and Information
Hohai University, HHU
Nanjing, CHINA
e-mail: lvxin.gs@163.com; njxufeng@163.com

Huaizhi Su
College of Water Conservancy and Hydropower Engineering
Hohai University, HHU
Nanjing, CHINA
e-mail: suhz@163.com

***Abstract*—*Ring signatures enable a user to sign a message so that a ring of possible signers is identified, without revealing exactly which member of that ring actually generated the signature. In some situations, however, an actual signer may possibly want to expose himself, for instance, if doing so, he will acquire an enormous benefit. In this paper, a signature scheme with designated verifiability based on Schnorr ring signature is proposed. The scheme provides a confirmation procedure, in which the real signer is able to convince a designated party that he is the one who generates the signature. The confirming procedure involves an interactive Zero-Knowledge proof protocol, which is non-transferable, and it only can be triggered by the signer. Based on the intractability of Discrete Logarithm Problem (DLP), the scheme is existentially unforgeable under adaptive-chosen message attack in the random oracle model.*

Keywords-Ring signature; Discrete logarithm problem; Designated verifiability; Unforgeability

1. INTRODUCTION

Digital signatures lay the foundations for identity authentication, data integrity, also non-repudiation, and it has been extensively used in the network society. Meanwhile, there is an imperative requirement of the ability to communicate anonymously for any privacy-preserving interactions in the applications on the Internet era, therefore, anonymous signature is gaining increasing attention recently. Ring signature is a kind of group-oriented anonymous signature technique. It allows the users to sign anonymously on behalf of a group (called ring) on his own choice, while ring members can be totally unaware of being conscripted in the ring. Any verifier can be convinced that a message has been signed by one of the members in this group, but the actual signer's identity is hidden. Unlike group signature[1], there is no group manager, thus the group formation of ring signature is spontaneous, or setup-free. Moreover, the anonymity of the signer cannot be revoked. Direct applications for ring signature include designated verifier signature[2] and secret leaking, but ring signature schemes are in general useful in applications where signer anonymity is desired.

2. RELATED WORK

The concept of Ring signature was firstly formalized by Rivest, Shamir and Tauman in 2001[3] on the background of "Leak Secret Securely", and they proposed an efficient ring signature scheme utilizing combing function and symmetric cryptosystem. Subsequently, there are continuous research works on ring signature in recent years[4-9].

Ring signatures provide an elegant way to leak authoritative secrets in an anonymous way, however, in some scenario the recipients want to give the leaker a big reward for his valuable intelligence (the secrets). Naturally everyone in the ring is eager to claim to be the signer of the ring signature. In this situation, a mechanism should be available for the actual signer to prove that he was the leaker. Lv et al.[10] firstly proposed a verifiable ring signature scheme in 2003, but it is complex for being based on double discrete logarithm. Lee et al.[11] proposed a convertible ring signature scheme, and it enables the real signer to turn their ring signature into an ordinary signature by releasing some information, thus any verifier can acquire his identity. Komano et al.[12] proposed a deniable ring signature scheme, allowing the verifier to identify who generates the signature through the interactions with the ring member. Klonowski et al.[13] advanced step-out ring signatures as an intermediate solution between the classical ring and group signatures. Dong et al.[14] extended Rivest's scheme[3] to a simple and efficient version in contrast to the available verifiable ring signatures. The glue value in the original scheme is replaced by the output of a hash function, and then the real signer is able to prove himself if he publishes the related secret value.

Because of being able to apply in many situations, verifiable ring signatures have become one of the most important branches of ring signature. In this paper, we propose a Schnorr ring signature scheme with designated verifiability (SRDV) which fits for the secret-leaking scenario. The scheme is to make the signer can prove the ownership of a ring signature at his will while satisfying the essential requirements of ring signature: unconditional anonymity and unforgeability. The confirmation procedure is a zero knowledge interactive proof with non-transferability, which means, the verifier cannot convince the others the actual identity of the real signer.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.48

3. PRELIMINARIES

3.1 Security Model

The scheme SRDV consists of the following algorithms (suppose there are n members in a ring):

- **Setup.** Given an unary string 1^k where k is a security parameter, the algorithm outputs a pair of public and secret keys (pk_i, sk_i) for each signer A_i and a list of system parameters **param** that includes k and the descriptions of a message space M as well as a signature space ψ and so on.
- **Sign.** Given the **param**, a message $m \in M$, the secret key sk_s of actual signer, and the public keys $(pk_1, pk_2, \ldots, pk_n)$ as input, the probabilistic algorithm outputs a ring signature $\sigma \in \psi$. In this process, two key parameters a^*, b^* (b^* is a power of a^* with an exponent v) are generated for confirmation, which will be involved in the signature σ.
- **Verify.** Given the message $m \in M$, and its signature $\sigma \in \psi$, also the public keys $(pk_1, pk_2, \ldots, pk_n)$ as input, the deterministic algorithm outputs **1** or **0** for **valid** or **invalid**, respectively. Before verifying the validity of the signature, there is a non-interactive proof performed by the verifier to check whether the key factors a^*, b^* are properly constructed without revealing any secret information. This make the verifier believe that someone (the actual signer) knows v, that is, the discrete logarithm of b^* to the base a^*.
- **Confirm.** It is a designated verifier zero-knowledge proof performed by the interaction between a prover and a designated verifier (DV) with input a^*, b^*, the verifier's public key and a set of parameters randomly selected in the procedure, making the prover convince the verifier that he is the actual signer who knows the secret v, namely the discrete logarithm of b^* to the base a^*.

3.2 Correctness and Security

Correctness. The scheme SRDV should satisfy the **verification correctness** -- signatures signed by honest signers are verified to be invalid with negligible probability, also the **confirmation correctness** -- only the actual signer can reveal himself and prove that he has created the signature.

Anonymity. The scheme is unconditional anonymous if for any set of n ring members, any message m and the corresponding signature σ, any adversary, even with unbounded computational power, cannot identify the actual signer with probability better than random guessing. That is, the adversary can only output the identity of the actual signer with probability no better than $1/n$.

Unforgeability. Any attacker must not have non-negligible probability of success in forging a valid ring signature for some message m on behalf of a ring that does not contain him, even if he knows valid ring signatures for messages, different from m, that he can adaptively choose.

Non-transferability. In SRDV, the DV cannot convince the others a certain ring member is the actual signer even if he had received his proof. That means if the DV transfers the proof to the others, the authenticity of the proof will be discredited.

4. SCHNORR RING SIGNATURE SCHEME WITH DESIGNATED VERIFIABILITY

In this section we present a novel verifiable ring signature scheme based on Schnorr ring signature[15].

SRDV-setup: Let p and q be large primes such that $q \mid p-1$ and $q \geq 2^k$, where k is the security parameter of the scheme. Let g, h are the generators of Z_p^* with order q, and let $H(\,)$ be a collision resistant hash function which outputs elements in Z_q, $H_1(\,)$ is a hash function: $\{0,1\}^* \to Z_p$, $F(\,)$ is a cryptographic hash function: $\{0,1\}^* \times \{0,1\}^* \to Z_q^*$.

Consider a set of potential signers $A_1, A_2, \ldots, A_n$. Each signer A_i has a private key $x_i \in Z_q^*$ and the corresponding public key $y_i = g^{x_i} \bmod p$.

SRDV-sign: To generate a designated verifiable ring signature on message m, the actual signer $A_s, s \in \{1, 2, \ldots, n\}$ executes the following steps:

1) Chooses $a, v \in Z_q^*$ at random and computes $a^* = h^a \bmod p, b^* = (a^*)^v \bmod p$.

2) Chooses $u \in Z_q^*$ at random and computes $U = (a^*)^u \bmod p, \theta = u + vF(m \| a^*, U) \bmod q$

3) A_s chooses $n-1$ random elements $a_i \in Z_q^*$ for $A_i (i \neq s)$, and computes $R_i = g^{a_i} \bmod p (i \neq s)$.

4) Computes $m' = H_1(m \| a^* \| b^*)$.

5) Chooses $a_s \in Z_q^*$ at random, and computes $R_s = g^{a_s} \prod_{i \neq s} y_i^{-H(m', R_i)} \bmod p$.

6) Computes $\sigma = \sum_{i=1}^{n} a_i + x_s H(m', R_s) \bmod q$.

7) The signature of the message m to be $(m, R_1, \ldots, R_n, \sigma, a^*, b^*, U, \theta)$.

SRDV-verify: given the message m and the corresponding ring signature, the verifier firstly checks whether the key factors a^*, b^* are properly constructed, and then verify the validity of the signature.

1) Checks whether $\left(a^*\right)^{\theta} = U\left(b^*\right)^{F\left(m\|a^*,U\right)} \bmod p$ holds or not.

2) Computes $m' = H_1\left(m \| a^* \| b^*\right)$ and $h_i = H\left(m', R_i\right)$ for all $1 \le i \le n$.

3) Checks if the equation $g^{\sigma} = R_1 \cdot R_2 \cdot \ldots \cdot R_n \cdot y_1^{h_1} \cdot y_2^{h_2} \ldots \cdot y_n^{h_n} \bmod p$ holds.

We can easily prove that if a^*, b^* has been well-constructed and the ring signature also has been legitimately generated, then the verification result is always **"1" (valid)**.

Remark 1. To prevent an attacker from using an existing pair of $\left(a^*, b^*\right)$ to sign a different message $\overline{m}$, m must be a portion of the input parameters of the function $F()$.

Remark 2. The whole signature is consisted of two parts: $\left(m, R_1, \ldots, R_n, \sigma\right)$ and $\left(a^*, b^*, U, \theta\right)$. The former one is the original Schnorr ring signature on m', and the latter one is Schnorr signature on $m \| a^*$. $\left(a^*, b^*\right)$ are taking part in the generation of σ, so these two parts are related.

SRDV-confirm: the actual signer convince the DV that he knows the discrete logarithm of b^* to the base a^* by using the following protocol. The protocol is an interactive proof, if the DV never loses his secret key, he can believe that the prover doesn't cheat him.

1) **Signer:** randomly chooses $\alpha \in_R Z_q^*$, and computes $\beta = \left(a^*\right)^{\alpha} \bmod p$; also chooses a random value $\gamma \in_R Z_q^*$, and computes the commitment $c = g^{\beta} PK_v^{\gamma} \bmod p$ to β, with $PK_v = g^{SK_v} \bmod p$ is DV's public key. Notice that PK_v is different from the public keys y_i used for signing. Then, the signer sends c to DV.

2) **DV:** randomly chooses $\varepsilon \in_R Z_q^*$ and sends it to the signer.

3) **Signer:** computes $\eta = \alpha + v\varepsilon \bmod q$ after receiving the challenge ε and sends it to DV, then the signer decommits c by revealing β and γ.

4) **DV:** checks whether the two equations $c = g^{\beta}\left(PK_v\right)^{\gamma} \bmod p$ and $\left(a^*\right)^{\eta} = \beta\left(b^*\right)^{\varepsilon} \bmod p$ hold. If they hold, DV confirms the prover is the actual signer of the message m.

5. SECURITY ANALYSIS

5.1 Anonymity

Theorem 1 The scheme SRDV proposed in Section 4 is unconditional anonymous.

Proof. Let $\left(m, R_1, \ldots, R_n, \sigma, a^*, b^*, U, \theta\right)$ be a valid ring signature of a message m. We can simply obtain the probability that A_s computes exactly the ring signature using SRDV:

$$\frac{1}{q-1} \cdot \frac{1}{q-2} \cdot \frac{1}{q-3} \cdot \frac{1}{q-1} \cdot \frac{1}{q-2} \cdot \ldots \cdot \frac{1}{q-n+1} \cdot \frac{1}{q-n}$$

which does not depend on A_s. Then the attacker outside the ring cannot distinguish the real signer from the others. Thus, the proposed scheme is unconditional anonymity. □

5.2 Unforgeability

Theorem 2 Under the security assumption of the Schnorr ring signature scheme and DLP (Discrete Logarithm Problem), the proposed scheme SRDV is secure against forgery in the random oracle model.

Proof. The theorem can be proved by contradiction, that is, if SRDV can be forged with a non-negligible probability, then DLP can be solved with a non-negligible advantage. For the detail, due to space limitation, you can find the similar proof in [16]. □

5.3 Non-transferability

Theorem 3 The proposed **SRDV-confirm** proof is non-transferable.

Proof. In the proposed interactive proof of SRDV-confirm, DV could be convinced of the proof if he recognizes that his secret key has not been compromised. However, if DV sends the "evidence" to any third party TP, the receiver can reasonably assume that the "evidence" may be false, since DV can use his secret key to forge a simulating transcript. That is, the TP is unable to distinguish a valid proof from an invalid proof forged by DV. The following steps demonstrate how DV fabricates a simulating transcript.

1) DV randomly selects $\delta \in Z_q^*$, and computes $c' = g^{\delta} \bmod p$, then sends it to the third party TP.

2) After receiving ε, DV decommits c' by revealing $\beta' = \left(a^*\right)^{\eta'}\left(b^*\right)^{-\varepsilon} \bmod p$, $\gamma' = \left(\delta - \beta'\right) \cdot \left(SK_v\right)^{-1} \bmod q$, with $\eta' \in Z_q^*$ is randomly chosen by DV.

3) TP checks the equations $c' = g^{\beta'}\left(PK_v\right)^{\gamma'} \bmod p$ and $\left(a^*\right)^{\eta'} = \beta'\left(b^*\right)^{\varepsilon} \bmod p$ whether hold or not.

Because TP cannot be sure whether the real signer performed the above transcript or not, he cannot actually be convinced of the proof. □

6. Efficiency Analysis

6.1 Space Complexity

Suppose the maximum length of the parameters and secret keys is L, and n denotes the number of ring member. The public keys of the system require $4L$; The public/private keys of ring members need $2nL$; In signing process, the actual signer needs $(n+6)L$ at most. So the requirement of storage space is $(3n+10)L$. Therefore, the proposed scheme is feasible in storage cost.

6.2 Computation Complexity

Let AD, MD, ED, H denote respectively modular addition, modular multiplication, modular exponentiation and hash operation. Amounts of calculation needed in **SRDV-sign**, **SRDV-verify** and **SRDV-confirm** are given in TABLE I.

TABLE I. Amounts of Calculation Needed in Each Process of The Proposed Scheme

Calculation	SRDV-sign	SRDV-verify	SRDV-confirm	
			Signer	DV
AD	2	0	1	0
MD	2	2	1	2
ED	2*n*+2	*n*+3	3	4
H	*n*+1	*n*+2	0	0

From the analysis in TABLE I, modular exponentiation and hash operation in **SRDV-sign** and **SRDV-verify** have the computation complexity of $O(n)$, owing to produce the values of the related parameters for all the ring member. However, the size of a ring is relatively stable, that is, the number of ring member is usually a constant. Therefore, the proposed scheme is feasible in computational efficiency.

7. Conclusion

In this paper, we firstly analyzed the current research progress on ring signature detailedly, and pointed out it is necessary to improve the available schemes along with the general applications. Aiming at a common scenario, a novel verifiable ring signature scheme was proposed (SRDV), based on an improvement of the original Schnorr ring signature scheme. SRDV has been proved to be unconditionally anonymous and existentially unforgeable in the random oracle model under the Discrete Logarithm assumption. The designated verifier is able to confirm the identity of the actual signer through 2 times interactions. The confirmation procedure is a zero-knowledge proof with non-transferability, making SRDV particularly suitable for many application environments.

Acknowledgment

This research is partially supported by "National Natural Science Foundation of China" (Grant No. 61272543); "National Key Technology Research and Development Program of the Ministry of Science and Technology of China" (Grant No. 2013BAB06B04); "Key Technology Project of China Huaneng Group" (Grant No. HNKJ13-H17-04); "Natural Science Foundation of Jiangsu Province" (Grant No. BK20130852); "Jiangsu Planned Projects for Postdoctoral Research Funds" (Grant No. 1401001C).

References

[1] D. Chaum and E. Heyst, "Group signatures," Advances in Cryptology - EUROCRYPT'91, LNCS 547. Berlin: Springer-Verlag, 1992, pp. 257-265.

[2] M. Jakobsson, K. Sako, and R. Impagliazzo, "Designated verifier proofs and their applications," Advances in Cryptology - EUROCRYPT'96, LNCS 1070. Berlin: Springer-Verlag, 1996, pp. 143-154.

[3] R. L. Rivest, A. Shamir, and Y. Tauman, "How to leak a secret," Advances in Cryptology - ASIACRYPT'01, LNCS 2248. Berlin: Springer-Verlag, 2001, pp. 552-565.

[4] J. F. Xiao, J. Liao, and G. H. Zeng, "Threshold ring signature for wireless sensor networks," Journal on Communications, 2012, vol. 33, no. 3, pp. 75-81 (in Chinese).

[5] A. J. Ge, C. G. Ma, and Z. F. Zhang, et al, "Identity-based ring signature scheme with constant size signatures," Chinese Journal of Computers, 2012, vol. 35, no. 9, pp. 1874-1880 (in Chinese).

[6] L. Z. Deng and J. W. Zeng, "Two new identity-based threshold ring signature schemes," Theoretical Computer Science, 2014, vol. 535, pp. 38-45.

[7] Y. Dodis, A. Kiayias, and A. Nicolosi, et al, "Anonymous identification in ad-hoc groups," Advances in Cryptology - EUROCRYPT'04, LNCS 3027. Berlin: Springer-Verlag, 2004, pp. 609-626.

[8] M. M. Tian, L. S. Huang, and W. Yang, "Efficient lattice-based ring signature scheme," Chinese Journal of Computers, 2012, vol. 35, no. 4, pp. 712-718 (in Chinese).

[9] P. P. Tsang and V. K. Wei, "Short linkable ring signatures for e-voting, e-cash and attestation," Information Security Practice and Experience - ISPEC 2005, LNCS 3439. Berlin: Springer-Verlag, 2005, pp. 48-60.

[10] J. Q. Lv and X. M. Wang, "Verifiable ring signature," DMS Proceedings. USA, 2003, pp. 663-665.

[11] K. C. Lee, H. Wei, and T. Hwang, "Convertible ring signature," IEE Proceedings, Communications, 2005, vol. 152, no. 4, pp. 411-414.

[12] Y. Komano, K. Ohta, and A. Shimbo, et al, "Toward the fair anonymous signatures: deniable ring signatures," Topics in Cryptology - CT-RSA 2006, LNCS 3860. Berlin: Springer-Verlag, 2006, pp. 174-191.

[13] M. Klonowski, Ł. Krzywiecki, and M. Kutyłowski, et al, "Step-out ring signatures," Mathematical Foundations of Computer Science 2008, LNCS 5162. Berlin: Springer-Verlag, 2008, pp. 431-442.

[14] Q. K. Dong, X. P. Li, and Y. M. Liu, "Two extensions of the ring signature scheme of Rivest-Shamir-Taumann," Information Sciences, 2012, vol. 188, pp. 338-345.

[15] J. Herranz and G. Saez, "Forking lemmas for ring signature scheme," Progress in Cryptology - INDOCRYPT 2003, LNCS 2904. Berlin: Springer-Verlag, 2003, pp. 266-279.

[16] X. Lv, Z. J. Wang, and F. Qian, et al, "Schnorr ring signature scheme with designated revocability," Intelligent Automation & Soft Computing, 2012, vol. 18, no. 6, pp. 739-749.

Improvement Research Based on affine encryption algorithm

Yongfeng Wu
School of Information,
Guizhou University of Finance and Economics
Guiyang,China
e-mail:wyf356@sina.com

***Abstract*—In the scientific development of today, people increasingly realize the importance of information security and secrecy. This greatly promoted the development of cryptology in information security, also it is social fields plays a important role. Therefore, cryptography to become important subject of security and confidentiality of communications.This paper compares traditional and modern cryptography,deeply analyses affine cryptography,and improves it by alphabet extension,block encryption and hash function,enhances the security and application of the algorithm.**

Keywords –cryptosystem*; *affine cipher*; *improved algorithm

I. INTRODUCTION

Cryptography focuses on the studying of secure communication,in order to through the changing of communication of information to prevent the interception from the third part.There are two independent branches of cryptography and cryptanalysis.Cryptography mainly study the encryption and the design of deciphering algorithm.On the other hand,Cryptanalysis focus on the analytical method to study the encrypted information,in an attempt to get the true ciphertext plaintext.

As a type of classic cryptography,the affine cipher has a long history,and the security and practicability of this cipher have been tested.The contribution of this paper lies in the analysis of affine cipher algorithm,and improves it through the mathematical theory to make it be a more practical cryptography algorithm.

II. TRADITIONAL ENCRYPTION ALGORITHM

A. Permutation cipher

The permutation cipher is to rearrange the letters in the text,the letter itself is unchanged,but its position has changed.

(1) Reverse permutation

The simplest permutation cipher is to reverse the order of the text,and then pile into a fixed length of the alphabet.

For example:

clear text:this cryptosystem is not secure.

After the permutation encryption,the length of each group is 4,then the cipher text is:eruc esto nsim etsy sotp yrcs iht.

(2) Matrix permutation

Another permutation cipher is to row the text into a matrix in a sequence,Then,in order to select the letters in the matrix to form the cipher text,The final cut into fixed length word line.

For example:

clear text:this cryptosystem is not secure.

Row matrix:

t	*h*	*i*	*s*	*c*	*r*
y	*p*	*t*	*o*	*s*	*y*
s	*t*	*e*	*m*	*i*	*s*
n	*o*	*t*	*s*	*e*	*c*
u	*r*	*e*			

Pick out the order:column

Ciphertext:tysnu hptor itete soms csie rysc

Above the first plaintext by row arrangement,according to read the column into ciphertext method,the actual equivalent to transpose the plaintext to ciphertext matrix.

B. Displacement substitution cipher

Displacement substitution cipher is a single table substitution is the most simple,In fact,each letter will be a bit forward to get the cipher text,Different can get different cipher text,The 26 letters are respectively corresponding to 0 – 25,As shown in the following table:

a	b	c	d	e	f	g	h	i
0	1	2	3	4	5	6	7	8
j	k	l	m	n	o	p	q	r
9	10	11	12	13	14	15	16	17
s	t	u	v	w	x	y	z	
18	19	20	21	22	23	24	25	

Encryption algorithm is as follows:

$E_k(i) = (i+k) \equiv j \bmod q, 0 \le i, j, k < q$;

For example:

If q=26,select k=5,There is the following transformation:

clear text:data security

Corresponding data sequence is:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.49

3 0 19 0 18 4 2 20 17 8 19 24
Ciphertext sequence:
8 5 24 5 23 9 7 25 22 13 24 3
Corresponding ciphertext:
i f y f x j h z w n y d

The number of key space elements q ,Where $k=0$ is an identity transform,In other words,The size of the key space depends on the size of the alphabet.The size of the alphabet is a constant,Therefore,Even using the approach taken to crack poor substitution displacement,Time encryption that is required is $O(C)$ constant level(Set a decryption time as a unit time).So this encryption algorithm has no use.

C. *Affine Cipher*

Affine cipher is one type of combination of shift cipher and multiplier password,and resulting in more options to get the key,the encryption algorithm is as follows:

$E_k(i)=ik_1+k_2\equiv j \bmod q$,

Condition: k_1 and q are the elements of each other,that is $(k_1,q)=1,0\le k<q$, $K_e=K_d=[k_1,k_2]$.

How to ensure $(k_1,q)=1$, Arbitrary k,m can not meet the conditions of k,m Coprime,But it can be constructed by the following method:

(1) k,m common denominator (k,m) obtained by Theorem;

(2) By theorem, $\left(\frac{k}{(k,m)},\frac{m}{(k,m)}\right)=1$,take

$k_1=\frac{k}{(k,m)}, q=\frac{m}{(k,m)}$,From any two numbers k,m ,and get a key coprime (k_1,q).

When $k_1=0$,Affine encryption algorithm on the shift encryption,When $k_2=0$,That multiplier encryption.

For example:take $k1=3,k2=6,q=26$.
Plaintext:a b c d e f g h
cipher text:g j m p s v y b

If this algorithm is used to encrypt the pure English character,Assumptions are case sensitive,But the capital remains encrypted ciphertext capital,Lowercase encrypted ciphertext still lowercase:

Take

$q=26,k_2\in[0,25]$ $and\, k_2\in N$,

$k_1\in\{1,3,5,7,9,11,15,17,19,21,23,25\}$,The secret key number in this case is: $26\times12-1=311$.

From the above discussion,although it could increase the cipher space from adopting this algorithm to encrypt,but the space is very limited.What's more the cipher can quickly be deciphered by even the Exhaustive method. Moreover,the affine cipher also has the similar backward of Replace cipher and shift cipher the mapping relationship between plaintext and ciphertext is fixed.Which means the same frequency analysis method can easily decipher the affine chipper.

III. THE IMPROVEMENT OF AFFINE ENCRYPTION ALGORITHM

A. *The extension of alphabet*

The affine encryption algorithm is

$E_k(i)=ik_1+k_2\equiv j \bmod q$,` k_1 and q are the elements of each other,that is $(k_1,q)=1,0\le k<q$, $K_e=K_d=[k_1,k_2]$.

According to the traditional method, $q=26$, the secret key number is 311.

We increase the value of q by enlarging the alphabet,in order to increase the total number of keys:

(1) Take $q=256$,the character corresponds to 8 bit ASCII;

(2) Take $q=256\times256=65536$,each character (string) uses double byte encoding,corresponding to the *Unicode* code.But the Unicode encoding is not necessarily just a double byte,it can be three bytes,four bytes.Therefore,the value of q can be greater according to the actual situation.

After the above improvement, $q=256,k_2\in[0,255]$ $and\, k_2\in N$, k_1 is all the odd numbers in the range of 256,the total key number is $256\times128-1=2^{15}-1=32767$,The improved key space is 105 times before improvement.If we take $q=256\times256=65536$,The total number will reach $2^{32}-1$.

B. *The packet encryption*

A clear set of text is divided into length L : $M=m_0,m_1,\cdots,m_L$, In order to expand the total number of keys,Each character m_i in the group is assigned an independent key K_i ,that is,the m 's key K is $2L$, $K=k_0,k_1,\cdots,k_{L-1},k_L,\cdots k_{2L-1}$,The corresponding key of m_i is $K_i=[k_i,k_{i+L}]$.

The security analysis are as followings:

(1) The same characters as the same group after the encryption is very likely to be the different charaters;

(2) The total number of keys is $2^{15\square L}-1$,the total

number of keys depends on the length of the key.

Theoretically,If using a key $q=256$ length of 16,The total number of keys for $2^{15\times 8}-1=2^{120}-1$,Provided the computing power of 2^{20} times/sec,The time required to decipher the 2^{75} years.

Actually we do not need such a long time,Because the average single table substitution algorithm can be deciphered by statistical methods.

Let n be the length of the plaintext L ,Encryption into ciphertext as follows:

$$M=\begin{pmatrix} m_{11} & m_{12} & \cdots m_{1L} \\ m_{21} & m_{22} & \cdots m_{2L} \\ \vdots & \vdots & \ddots \ \vdots \\ m_{n1} & m_{n2} & \cdots m_{nL} \end{pmatrix} \text{Encryption} \rightarrow C=\begin{pmatrix} c_{11} & c_{12} & \cdots c_{1L} \\ c_{21} & c_{22} & \cdots c_{2L} \\ \vdots & \vdots & \ddots \ \vdots \\ c_{n1} & c_{n2} & \cdots c_{nL} \end{pmatrix}$$

Decipher analysis:

The above matrix decomposition by column(Each column is a vector):

$M=(\alpha_1,\alpha_2,\cdots,\alpha_L)$,

Among those, $\alpha_i=(m_{1i},m_{2i,}\cdots,m_{ni})^T, 1\le i\le L$;

$C=(\beta_1,\beta_2,\cdots,\beta_L)$,

Among those, $\beta_i=(c_{1i},c_{2i,}\cdots,c_{ni})^T$, $1\le i\le L$;

The last to know, α_i in any one of the elements use the same key $K_i=[k_i,k_{i+L}]$.

As long as the algorithm is clear,it is easy to be able to get the text in the frequency classification and decoding.Therefore,the security of the modified affine encryption algorithm is not high,it is necessary to further study the practical application of the proposed method.

C. Hash algorithm

After B of the improved affine encryption,although there are a lot of key space,but the actual is still not safe.So we use a function to hash.

Using the *Hash* function: $Hash=i\square k_1+k_2\equiv j \bmod len$,

$$\text{Condition:}\begin{cases}(k_1,len)=1;\\ 0\le k_1,k_2<len\end{cases}$$

Obviously,the hash is actually a set of address sequence affine cipher,as the new address.

Hypothesis:

Plaintext $M=(m_1,m_2,\cdots,m_i,\cdots,m_n)$;

Auxiliary matrix $A=(a_1,a_2,\cdots,a_i,\cdots,a_n)$;

cipher text $C=(c_1,c_2,\cdots,c_i,\cdots,c_n)$;

M,A,C the length of the three matrices are len

Affine encryption Improved,Not enter C to encrypt M ,Instead,the encrypted text is entered into the auxiliary matrix A ,After the address of the HASH A as the C to store the address of the encrypted text,This will use the same key ciphertext hash.

D. Secret key construction

Analysis of the improved security before affine Cipher:

(1) The encryption process has linear characteristics,Hashing process has linear characteristics;

(2) If part of the key is deciphered,it will lead to some ciphertext is deciphered,That key is not long enough (good key and plaintext to be approximately equal in length) and the lack of randomness.

Over the first 1 linear characteristic can be reduced by several rounds of encryption,hashing,Therefore,we can use some specific ways to continue to construct a new key.

Construction method:Set up the first round key for $K_0=k_1,k_2,\cdots k_n$,

(1)It generates two random offset array

D_{01}, D_{02} ($D_{0i}=Hash(\{0,1,\cdots n\})$,

(2) $S_0=[K_0/2]-D_{01}+q \bmod q$,

(3) $T_0=[K_0/2]+D_{02} \bmod q$,

(4) $R_0=Hash(K_0)$,

(5) $K_1=S_0\square T_0+R_0 \bmod q$,

Repeated execution (1) - (5) can continue to construct a new key,So that keys can be approximated as long as the plaintext.

IV. COMPREHENSIVE IMPROVEMENT ALGORITHM

1) Symbol description:

(1)Plaintext M is divided into groups of length L plaintext,that is $M=M_0M_1$,Among them $M_i=m_{i0}m_{i1}\cdots m_{i(L-1)}$,The cipher group is $C=C_0C_1\cdots, C_i=c_{i0}c_{i1}\cdots c_{i(L-1)}$,the key number is $K=K_1K_2\cdots$,

$K_i=K_i^{'}K_i^{''}=k_{i0}k_{i1}\cdots k_{i(L-1)},k_{iL}k_{i(L+1)}\cdots k_{i(2L-1)}$,Assu

me that M, C and K have been converted to the corresponding encoding;

(2)The actual K_i length may be $2L$ or $2L+1$,Therefore,the length of K_i is len;

(3)The auxiliary matrix $A = A_0 A_1 \cdots$ is used to store the intermediate text in the process of encryption;

(4) $hash$ function is defined as: $hash(array, parameter1, parameter2)$,Which L represents an array of hash, $parameter1$ *and* $parameter2$ are hashed parameter;

(5)Setting the security level:Lower(security=1),Intermediate(security=2),Senior(security=3),Special grade(security=4);

(6) P_i is an array of parameters in the i round, p_{ij} is one of the elements, $hp1, hp2$ is the hash parameter of the whole Ciphertext;.

(7) D_1, D_2 is the offset parameter array, $D_0 = \{0,1,\cdots,len-1\}$;

2) Specific encryption steps are as follows:($i=0, j=-1$)

(1) $P_i = K_i \bmod len$;

(2) $hp1 = hp1 + p_{i(++j \bmod len)}$, $hp2 = hp2 + p_{i(++j \bmod len)}$;

(3) $D_{i1} = hash(D_0, p_{i(++j \bmod len)}, p_{i(++j \bmod len)})$, $D_{i2} = hash(D_0, p_{i(++j \bmod len)}, p_{i(++j \bmod len)})$;

(4)Affine cryptography: $E_{K_i}(M_i) = M_i K_i^{'} + K_i^{''} \equiv A_i \bmod q$;

(5) $C_i = H(A_i, p_{i(++j \bmod len)}, p_{i(++j \bmod len)})$;

(6)If $(i+1)*L < C.len$,So the next step,Otherwise,step 9;

(7)

$$S_{i+1} = H(([K_i/2] - D_{i+1} + q) \bmod q, p_{i(++j \bmod len)}, p_{i(++j \bmod len)})$$

$$T_{i+1} = H(([K_i/2] + D_{i+1}^{'}) \bmod q, p_{i(++j \bmod len)}, p_{i(++j \bmod len)})$$

$$R_{i+1} = H(K_i, p_{i(++j \bmod len)}, p_{i(++j \bmod len)});$$

(8)

$K_{i+1} \equiv S_{i+1} T_{i+1} + R_{i+1} \bmod q$, $i = i+1$,Returning to step 1;

(9) $C = H(M, hp1, hp2)$;

security=security-1;

(10)If security>0 ,then $i=0$,Go back to Step 1,Otherwise encryption complete.

V. CONCLUSIONS

The paper introduce the classic encryption and modern encryption,particularly analysis the affine cipher,and improve the practicability and security of affine cipher through the algorisms improvement including extended alphabet,packet encryption,hashing address.

REFERENCES

[1] Chunfu Jia,"Information security mathematics foundation", Beijing:Tsinghua University press,pp.59-64.2010.

[2] Mingquan Zhou,Lintao Lv and Junhuai Li,"Network information security technology".Xi'an:Xi'an electronic publishing house,pp.102-131.2010.

[3] Wenchang Shi,"Introduction to security of information systems (Second Edition)".Beijing:Publishing House of Electronics indusy,pp.114-128.2014.

[4] (Canada) Stimson,"Cryptography Principles and Practice (3rd Edition)".beijing:Publishing House of Electronics indusy,pp.135-152.(In Chinese),2009.

[5] Xueguang Zhou,Yi Liu,"Information Security Studies".beijing:Machinery Industry Press,pp.56-88.(In Chinese),2003

[6] W illian Stallings,"Cryptography and Network Security:Principles and Practice".Beijing:Publishing House of Electronics indusy,pp.2021-226.(In Chinese),2003.

[7] Bo Yang,"Modern cryptography". beijing :Tsinghua University press,pp.69-102.2007.

[8] Bruce Schneier,"Applied Cryptography:Protocols,Algorithms and C source code".beijing:Machinery Industry Press,pp.79-92.2014.

[9] (US) Stallings,"Cryptography and network security - principle and practice of encoding".Beijing:China Water&Power Press),pp.79-92.2012.

System design and obstacle avoidance algorithm research of vacuum cleaning robot

Li Guangling/ Jiangyin Polytechnic College
Jiangsu Engineering R&D Center for Information Fusion Software
Wuxi, China
lglzjm@163.com

Pan Yonghui/ Jiangyin Polytechnic College
Jiangsu Engineering R&D Center for Information Fusion Software
Wuxi, China
pyh828@sina.cn

***Abstract*—Designing multi-sensor hardware system based on Cortex-M0 controller, detecting the surrounding of vacuum cleaning robot by sensor combination of ultrasonic, infrared, collision, etc., we proposed an information fusion algorithm with Kernel PCA based on SFC neural network to process the multi-sensor data. And the output result of SFC neural network is used to control the localization and obstacle avoidance of vacuum cleaning robot. The experimental results demonstrate that the multi-sensor hardware system and obstacle avoidance algorithm based on SFC neural network proposed in this paper can improve the localization and obstacle avoidance accuracy of vacuum cleaning robot highly, which is robust for different working environment of vacuum cleaning robot.**

Keywords-vacuum cleaning robot; system design; Kernel PCA; supervised fuzzy clustering; SFC neural network

Ⅰ. INTRODUCTION

Vacuum cleaning robots, or home service robots can clean room automatically. The robot, integrated with mechanics, electronics, sensing, controlling, computer technology and artificial intelligence can clean the room on its way by using its small dust-vacuum component. To make cleaning robot work, R&D staff must monitor robot position, attitude, speed and system internal status, perceive static and dynamic information around the robot environment to make vacuum cleaning robot operate smoothly and dynamically adapt to the working environment changes, and use the multi-sensor information fusion technology to achieve the robot positioning, obstacle avoidance and environmental modeling [1]. The purpose of the robot location is to tell its position as well as a variety of objects, and navigation is intended to guide the robot to reach the target position. Obstacle avoidance, occurring in the procedure of robot location and navigation, is an important manifestation of robot autonomy [2-3].

With the recent development of battery storage technology, domestic and foreign automatic vacuum cleaning robot has witnessed a major breakthrough, prestigious universities and electrical companies have launched a home service robot products with its own brand. Despite the great achievements in research and development of vacuum cleaning robots home and abroad, many key technologies, especially environmental awareness and obstacle avoidance in the variational environments need to be further improved.

With the multi-sensor hardware system design in mind, this paper proposed obstacle avoidance algorithm based on supervised fuzzy clustering (SFC) neural network.

Ⅱ. ROBOT HARDWARE SYSTEM DESIGN

A. *Travelling mechanism design*

Chassis of vacuum cleaning Robot is made from lighter porous circular aluminum which has a certain strength. It equipped with four wheels, with left and right wheels as its driving wheels, front and back wheels as its driven wheels. Power supply is high current rechargeable lithium battery pack with 15 V and 6 V dual output voltage, can provide power to the stepper motor and system control circuitry respectively. Left and right driving wheels are controlled by stepper motor to make robot move and steer. Robot speed control model is shown in Figure 1.

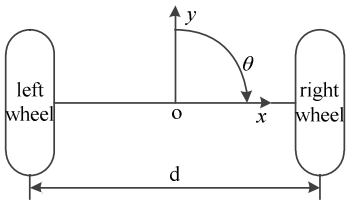


Figure 1. Module of robot speed control

In Figure 1, Point O is a Cartesian coordinate center, d is the distance between the two driving wheels, θ is the direction angle of the robot. Assume that there is no sliding occurred during motion, the relationship between the instantaneous speed of y, x, θ directions and that of left and right wheel speed is:

$$y = \frac{v_l + v_r}{2}, x = 0 \quad (1)$$

$$\theta = \frac{v_l + v_r}{d} \quad (2)$$

v_l and v_r in formula (1) and (2) represent left and right moving speed respectively.

B. *Circuit Design*

Vacuum cleaning robot system uses Cortex-M0 (LPC1114) as its core control center, including power supplies, motor driving, liquid crystal display and various sensors. Control circuit diagram is shown in Figure 2.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.50

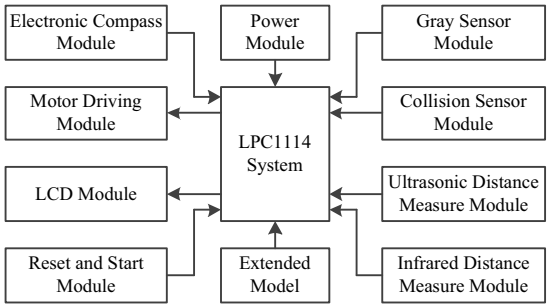


Figure 2. Frame of system control circuit

III. KERNEL PRINCIPAL COMPONENT ANALYSIS

In this paper, we first use correlation analysis method to get correlation coefficients among the multi-sensor information data. Then we remove those indexes which are highly correlated with each other and those scarcely related to motion control of robot, and apply kernel PCA method to analyze the reserved indexes to get rid of the redundant information [4].

Let us first start with a set of M centered data, $\mathbf{x}_k$, in the input space, $k = 1, \cdots, M$, $\mathbf{x}_k \in \mathbf{R}^N$. Linear principal component analysis requires the diagonalization of M-sample estimate of the covariance matrix

$$\mathbf{C} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_i \mathbf{x}_i^T \tag{3}$$

With an intent to find eigenvalues ($\lambda \geq 0$) and the associated eigenvectors $\mathbf{v}$ satisfying,

$$\lambda \mathbf{v} = \mathbf{C}\mathbf{v} \tag{4}$$

It would be useful to note that for non-negative eigenvalues, all solutions must lie in the span of input data. Thus the eigenvalue equation can be written as

$$\lambda(\mathbf{x}_k \cdot \mathbf{v}) = (\mathbf{x}_k \cdot \mathbf{C}\mathbf{v}),\ k = 1, \cdots, M \tag{5}$$

For KPCA, we first define a nonlinear mapping of the centered input data in the feature space as $\Phi : \mathbf{R}^N \rightarrow F$.

The problem can now be formulated as the diagonalization of the M-sample estimate of the covariance matrix in the high dimensional feature spaces [5].

$$\mathbf{C} = \frac{1}{M}\sum_{i=1}^{M} \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T \tag{6}$$

Where $\Phi(\mathbf{x}_i)$ are centered nonlinear mapping of the input variables. Here, we need to find the nonnegative eigenvalues λ and eigenvectors $\mathbf{v}$, satisfying the equation

$$\lambda \mathbf{v} = \hat{\mathbf{C}}\mathbf{v} \tag{7}$$

Noting that all the eigenvalues lie in the span of the transformed data in the high dimensional space, the equivalent relation can be written as

$$\lambda(\Phi(\mathbf{x}_k) \cdot \mathbf{v}) = (\Phi(\mathbf{x}_k) \cdot \mathbf{C}\mathbf{v}),\ k = 1, \cdots, M \tag{8}$$

Also, the coefficients α can be related to $\mathbf{v}$ as

$$\mathbf{v} = \sum_{i=1}^{M} \alpha_i \Phi(\mathbf{x}_i) \tag{9}$$

Combination of fomula (6), (8) and (9) yields

$$\begin{aligned} &\lambda \sum_{i=1}^{M} \alpha_i (\Phi(\mathbf{x}_k) \cdot \Phi(\mathbf{x}_i)) \\ &= \frac{1}{M}\sum_{i=1}^{M} \alpha_i \left(\Phi(\mathbf{x}_k) \cdot \sum_{j=1}^{M} \Phi(\mathbf{x}_j)\right)(\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)) \\ &\qquad \forall k = 1, \cdots, M \end{aligned} \tag{10}$$

Further we dafine an M×M kernel matrix K such that

$$K_{ij} = (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) \tag{11}$$

KPCA makes use of the fact that an inner product in the feature space has an equivalent kernel in the input space. Thus it is neither necessary to know the form of the function, $\Phi(\mathbf{x})$ nor we need to calculate the dot product in the very high dimensional space. We can thus employ appropriate kernels to evaluate this in the input space itself.

IV. SFC-RBF NEURAL NETWORK

A. RBF Neural Network

The topology of the RBFNN is shown in figure 3.

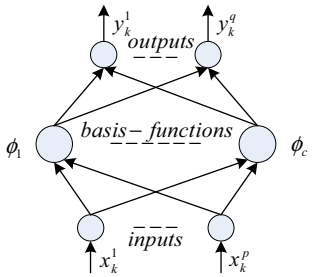


Figure 3. Topology of RBFNN

Each hidden node evaluates a kernel function (receptive field) $\phi_i(\mathbf{x})$ on the incoming input, and the output $y(\mathbf{x})$ is simply a weighted linear summation of the output of the kernel functions:

$$y(\mathbf{x}) = \sum_{i=0}^{c} w_i \phi_i(\mathbf{x}) \tag{12}$$

In the case of the Gaussian basis functions, for example, we have

$$\phi_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{v}_i\|^2}{2\sigma^2}\right) \tag{13}$$

In our network, the RBF kernels do not assume any explicit functional form such as Gaussian, ellipsoidal, etc., but directly rely on the computation of the relevant distances. Let us suppose to have already determined the kernel function centers $\mathbf{v}_1, \cdots, \mathbf{v}_c$. Let us denote the obtained levels of matching by $m_1, m_2, \cdots, m_c$. The matching level m_i is inversely proportional to the distance between $\mathbf{x}$ and the prototype of the ith RBF unit, $\mathbf{v}_i$. Since these levels sum up to one (for proper normalization), this leads us to the optimization problem

$$\min_{m_1, \cdots, m_c}\left\{\sum_{i=1}^{c} m_i^2 \|\mathbf{x} - \mathbf{v}_i\|^2\right\} \tag{14}$$

Subject to

$$\sum_{i=1}^{c} m_i = 1 \tag{15}$$

The use of the Lagrange multipliers method leads to the solution

$$m_i = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|\mathbf{x}-\mathbf{v}_i\|^2}{\|\mathbf{x}-\mathbf{v}_j\|^2}\right)} = \phi_i(\mathbf{x}) \tag{16}$$

We shall see how to properly modify this expression in order to fully exploit the fuzziness involved in the procedure to determine the basis function centers; for now it is enough to say that the neuron situated in the output layer carries out a linear combination of the matching levels, yielding

$$y_k^i = \sum_{i=1}^{c} w_i m_i \tag{17}$$

Where $w_1, w_2, \cdots, w_c$ are the hidden-to-output weights. This expression can be formulated in a matrix notation as follows:

$$\mathbf{y}(\mathbf{x}) = W\Phi \tag{18}$$

Where $W = (w_{ij})$ and $\Phi = (\Phi_{ij})$. We can optimize the weights by minimization of a suitable error function. It is particularly convenient to consider a sum-of-squares error function given by

$$E = \frac{1}{2}\sum_{n}\left|\mathbf{y}^n(\mathbf{x}) - \mathbf{t}^n\right|^2 \tag{19}$$

Where $\mathbf{t}$ is the target value for the output unit when the network is presented with input vector $\mathbf{x}$. Since the error function is a quadratic function of the weights, its minimum can be found in terms of the solution of a set of linear equations:

$$\Phi^T\Phi W^T = \Phi^T T \tag{20}$$

Where $(T) = \mathbf{t}$ and $(\Phi) = \phi(\mathbf{x})$. The formal solution to the weights is given by

$$W^T = \Phi^P T \tag{21}$$

Where the notation Φ^P denotes the pseudo-inverse of Φ. Thus, the second layer weights can be found by fast, linear matrix inversion techniques.

B. Fuzzy C-Means Clustering

The Fuzzy c-means (FCM) clustering algorithm is a set-partitioning method based on Picard iteration through necessary conditions for optimizing a weighted sum of squared errors objective function J_m.

The FCM algorithm was developed to minimize the objective function

$$J_m = \sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m d(\mathbf{x}_k, \mathbf{v}_i),\ 1 < m < \infty \tag{22}$$

In formula (22), $d(\mathbf{x}_k, \mathbf{v}_i)$ is any inner product norm metric of the distance between the feature vector $\mathbf{x}_k \in X$ and the prototype $\mathbf{v}_i \in R^n$. A metric often used in applications is the squared Euclidean distance between $\mathbf{x}_k$ and $\mathbf{v}_i$, that is $d(\mathbf{x}_k, \mathbf{v}_i) = \|\mathbf{x}_k - \mathbf{v}_i\|^2$. The coupled first order necessary conditions for solutions (U, V) to $\min\{J_m(U, V)\}$ are

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|}\right)^{(2/(m-1))}},\quad \begin{matrix}1 \le i \le c\\ 1 \le k \le N\end{matrix} \tag{23}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^m},\quad 1 \le i \le c \tag{24}$$

C. Supervised Fuzzy C-Means Clustering

In this section, we extend the original FCM objective function used by linear summation sub-networks and propose a supervised fuzzy c-means clustering (SFCM) model [6-8]. Rather than defining J based on $\mathbf{x}_k \in R^n$ only, we supply it with information on the output space by defining a new objective function which assumes the following form

$$J = \sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m\left(\|\mathbf{x}_k - \mathbf{v}_i\|^2 + \left|\mathbf{y}_k - \mathbf{y}_k^*\right|^2\right) \tag{25}$$

Where $\mathbf{y}_k$ and $\mathbf{y}_k^*$ are the corresponding desired output and computing output of sub-networks respectively. The first term of formula (25) denotes fuzzy c-partitions of input patterns $\mathbf{x}_k$ by minimizing the distance between inputs $\mathbf{x}_k$ and prototypes $\mathbf{v}_i$. The second term of formula (25) requests the computing output of system to approach the desired output mostly.

Now, by applying the Lagrange multipliers technique to formula (25), we derive the necessary conditions for the partition matrix and the prototypes, namely

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|^2 + \left|\mathbf{y}_k - \mathbf{y}_k^*\right|^2}{\|\mathbf{x}_k - \mathbf{v}_j\|^2 + \left|\mathbf{y}_k - \mathbf{y}_k^*\right|^2}\right)^{1/(m-1)}},\quad 1 \le i \le c, 1 \le k \le N \tag{26}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N}(u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{N}(u_{ik})^m},\ 1 \le i \le c \tag{27}$$

The structure of SFCM is shown in figure 4.

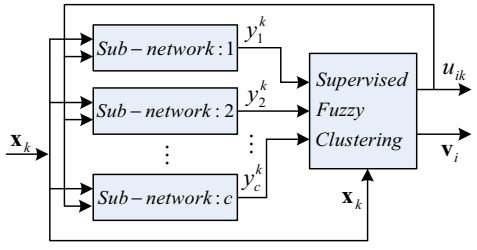


Figure 4. Construction of SFCM

There are two parts in SFCM model. One is supervised fuzzy classifier, and the other is linear summation sub-networks [9].

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \subset R^n$ be the patterns to cluster and $\mathbf{y}_k \in R^q, k = 1, \cdots, N$ be the corresponding desired output. Suppose to organize the data in c clusters, we can associate a local linear summation model for each cluster [10]. The topology of the linear summation sub-network is shown in figure 5.

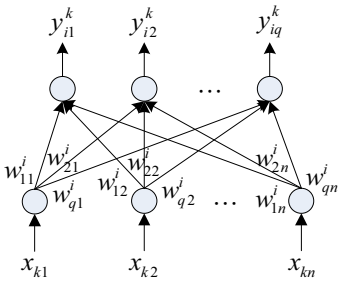


Figure 5. Topology of sub-network

Where $\mathbf{w}_i = (\mathbf{w}_1^i, \mathbf{w}_2^i, \cdots, \mathbf{w}_q^i)^T$, $\mathbf{w}_{ij} = (w_{j0}^i, w_{j1}^i, \cdots, w_{jn}^i)$, and the output of ith sub-network is $\mathbf{y}_k = \mathbf{w}_i \mathbf{x}_k$.

In order to calculate the weights of the ith sub-network $\mathbf{w}_i \in R^n$, $i = 1, \cdots, c$, we solve c least square problems, that is, one for each sub-network:

$$\min V = \frac{1}{2} \sum_{k=1}^{N} \{u_{ik} \mathbf{w}_i \mathbf{x}_k - \mathbf{y}_k\}^2 \tag{28}$$

We can rewrite this expression in the form

$$\min V = \frac{1}{2} \sum_{k=1}^{N} \{\mathbf{w}_i \psi_i(\mathbf{x}_k) - \mathbf{y}_k\}^2 \tag{29}$$

Where $\psi_i(\mathbf{x}_k) = u_{ik} \mathbf{x}_k$.

Differentiating formula (29) with respect to the parameters $\mathbf{w}_i$ and setting the derivative to zero we obtain

$$\frac{\partial V}{\partial \mathbf{w}_i} = \sum \{\mathbf{w}_j \psi_j(\mathbf{x}_k) - \mathbf{y}_k\} \psi_i(\mathbf{x}_k) = 0 \tag{30}$$

Where $j = 1, \cdots, c$, Writing formula (30) in matrix notation we have $(\psi^T \psi) W^T = \psi^T Y$, and then

$$W^T = \psi^P Y \tag{31}$$

Where ψ^P denote the pseudo-inverse of matrices ψ.

Ⅴ. Experimental Results

A. KPCA Analysis of Multi-sensor Information

Now, let us extract principal components of multi-sensor data of the vacuum cleaning robots using kernel PCA method. And the anterior ten kernel principal components are shown in figure 6. And the accumulative contribution ratio of anterior six kernel principal components is 91.2%.

The definitions of anterior six kernel principal components are shown in Table 1.

TABLE 1. DEFINITION OF KERNEL PCS

Kernel PC	Perceiving Index	Relative Sensors
KPC1	Collision Connection	Collision Sensor
KPC2	Front Distance	Ultrasonic and Infrared Sensor
KPC3	Left Distance	Ultrasonic and Infrared Sensor
KPC4	Right Distance	Ultrasonic and Infrared Sensor
KPC5	SideStep	Ultrasonic Sensor
KPC6	Direction Angle	Electronic Compass Sensor

B. Simulation of SFCNN Algorithm

Different simulation environments are applied to test the location and obstacle avoidance algorithm of vacuum cleaning robots'. The position and number of platform, baffle, obstacle in the simulation environment are set randomly. It requires vacuum cleaning robot, spending as little time as possible to cover the largest space floor, and do not fall from the platform.

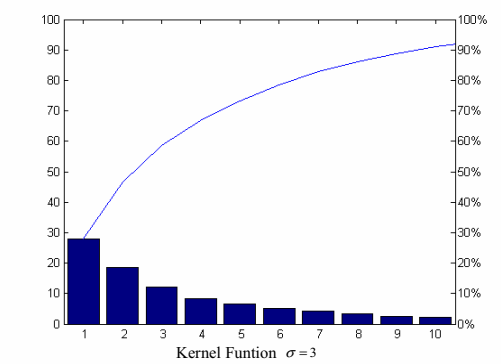


Figure 6. KPCA contribution of multi-sensor data

The extracted six kernel principal components and two driving wheels controlling of robot are as the input and the output of neural network respectively. Choose 6 input layer nodes of supervised fuzzy clustering neural network, each node corresponding to perceiving index and relative sensors are shown in Table 1. Choose 12 nodes as hidden layer (fuzzy layer and rules computing layer) and 2 output nodes to control the moving speed of the left and right driving wheels of vacuum cleaning robot.

Getting a large number of information data of working environment and status based on multi-sensor in the course of the actual travelling of the robot, normalized data and initialization parameters of supervised fuzzy neural network, We uses Matlab to train SFC neural network. By large numbers of testing, we find the convergence speed of our algorithm is very fast on the sum-square error 0.1, and the iteration number of SFCM is about 10^3 degree. Using 3 groups of standardized data (each group has 20 sample points) for algorithm simulation, it can be seen that SFC neural network system overall prediction accuracy is very good. Among 60 sample points, the maximum absolute error

is less than 0.01, and the maximum relative error is less than 3.6%. The test results are shown in Figure 7.

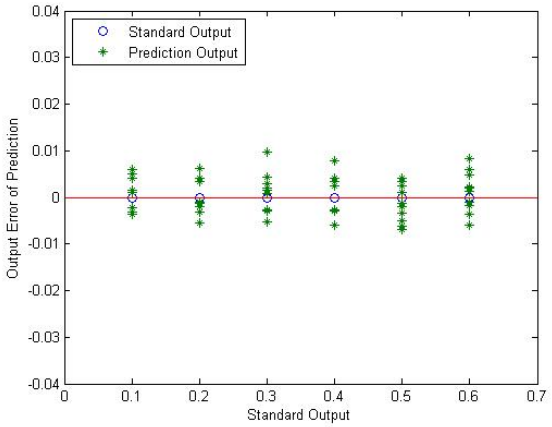


Figure 7. Simulation of SFCNN

The trained algorithm was transplanted into the robot controlling chip and running it, by changing the layouts of the simulation environment many times to test the robot, we found that the robot motion control accuracy is above 93%, and obstacle avoidance rate close to 100%.

VI. CONCLUSIONS

This paper uses the multi-sensor information fusion algorithm based on supervised fuzzy clusterin neural network to process the multi-sensor data of the vacuum cleaning robots. The output result of SFCNN is used to control the robot motion.

After repeated experiments, it shows that the algorithm has better adaptability in different working environment. In the future, we will continue to optimize the sensor assembly and distribution of vacuum cleaning robot, and improve the signal pre-processing circuit in order to better improve the robot operational control accuracy and the robustness of obstacle avoidance algorithm.

ACKNOWLEDGMENT

This work was financially supported by the Opening Project of Jiangsu Engineering R&D Center for Information Fusion Software (SR-2013-03), sponsored by the Project of Jiangsu Province Natural Fund (BK2012128).

REFERENCES

[1] B. Dorj, D.Tuvshinjargal, K. Chong, D. Hong and D. Lee, "Multi-Sensor Fusion Based Effective Obstacle Avoidance and Path-Following Technology," Advanced Science Letters, vol. 20, No. 10-12, 2014, pp. 1751-1756, doi:10.1166/asl.2014.5680.

[2] A. Reyaz, B. Baasandorj, S.Ho-Park, D. Lee and K. Chong, "Mobile Robot Obstacle Avoidance Techniques," Advanced Science Letters, vol. 20, No. 10-12, 2014, pp. 1927-1931, doi:10.1166/asl.2014.5683.

[3] K. Charalampous, I. Kostavelis, A. Amanatiadis and A. Gasteratos, "Real-Time Robot Path Planning for Dynamic Obstacle Avoidance," Journal of Cellular Automata, vol. 9, Mar. 2014, pp. 195-208.

[4] K. Jorgensen and L. Hansen, "Model Selection for Gaussian Kernel PCA Denoising," IEEE Transactions on Neural Networks and Learning Systems, vol. 23, No. 1, Jan. 2012, pp. 163-168, doi:10.1109/tnnls.2011.2178325.

[5] A. Jade, B.Srikanth, V. Jayaraman and L. Priya, "Feature Extraction and Denoising Using Kernel PCA," Chemical Engineering Science, vol. 58, 2003, pp. 4441-4448.

[6] B. Hartmann, O. Banfer, and I. Skrjanc, "Supervised Hierarchical Clustering in Fuzzy Model Identification," IEEE Transactions on Fuzzy Systems, vol. 19, No. 6, Dec. 2011, pp. 1163-1175.

[7] G. Castellano, A. Fanelli, and M. Torsello, "Shape Annotation by Semi-Supervised Fuzzy Clustering," Information Sciences, vol. 289, Aug. 2014, pp. 148-161.

[8] G. Donghai, Y. Weiwei, L.Young-Koo, G. Andrey and L. Sungyoung, "Improving Supervised Learning Performance by Using Fuzzy Clustering Method to Select Training Data," Journal of Intelligent & Fuzzy Stems, vol. 19, 2008, pp. 321-334.

[9] A. Staiano, R. Tagliaferri and W. Perycz, "Improving RBF Networks Performance in Regression Tasks by Means of a Supervised Fuzzy Clustering," Neurocomputing, vol. 69, 2006, pp. 1570-1581.

[10] L. Wen and Y. Hori, "An Algorithm for Extracting Fuzzy Rules Based on RBF Neural Network," IEEE Transactions on Industrial Electronics, vol. 53, No. 4, Aug. 2006, pp. 1269-1276.

Software Quality Issues and Challenges of Internet of Things

Dr. Jay Kiruthika
Kingston University
jay.kiruthika@gmail.com

Dr. Souheil Khaddaj
Kingston University
s.khaddaj@kingston.ac.uk

***Abstract*— Internet of things (IoT) is making its mark on various aspects of life. Determining the quality features for such devices vary according to the functionalities of the system. Non-functional quality factors play a vital role in evaluating such systems due to their applicability and multiple functionalities at a given time. Thus, this paper focuses on addressing quality of service (QoS) issues and considers the crucial factors when designing quality models for IoT systems and the challenges that need to be addressed.**

Keywords- Internet of Things (IoT); Quality of service (QoS); Quality Modelling;

I. INTRODUCTION

Internet of things or IoT as it is referred to, is invading all aspects of our live. Its agility to perform multiple functions and ability to connect to various devices has brought in complex coding and interconnectivity to various systems. The push/pull algorithms have to be tuned for specific functions as they have to be personalized to specific purposes. Some of the IoT systems use location services to determine or to perform tasks. The use of sensors for intelligent decision making in IoT systems, to improve user specific functionalities, are complex by nature. IoT is a compendious utilization of existing technologies via new communication modes. It blends various technologies together like the pervasive networks, miniaturization of devices, mobile communication, middleware components and end user devices.

The future IoT devices will have increase access to the data coming from internet-enabling manufacturing process, workers, assets, components, exchanging information, finished products and clients etc. forcing the interconnectivity aspect of the system as the primary process to focus on. Moreover, the non-functional quality aspects such as performance, usability, timelines, correctness, security needed to be tuned and should be traceable for the IoT systems to function efficiently especially if they are based on Cloud solutions and handling big data.

Furthermore, IoT's machine to machine connectivity can be used to track smart devices as well as simple products connected via RFID tags. Companies like ORBCOMM a satellite communications and remote-measurement Company use such systems for sensor-enabled notification system for asset tracking. As with the rest of all the emerging technology, IoT has spread its wings to various industries [1] like smart agriculture, logistics, mobile telephony, intelligent transportation, smart grid, smart environmental protection, smart safety, smart medical, smart home etc. and the future possibilities are endless (Figure 1). The maintaining of such systems and challenges IoT faces now and in the near future are discussed next.

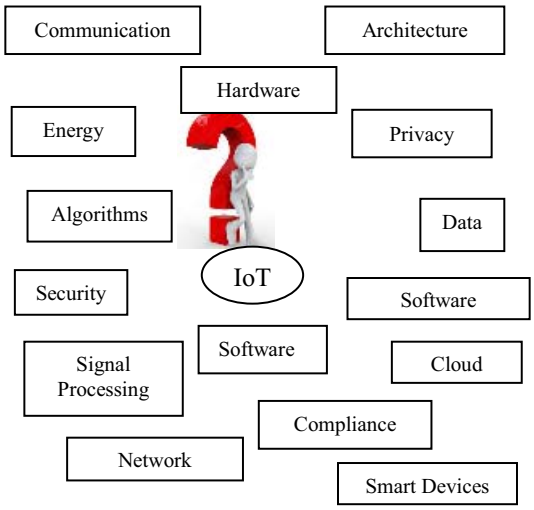


Figure 1. Challenges of IoT

In this work factors in software quality that should be taken into account when using an IoT based system will be considered with the aim of building a QoS model for Internet of Things.

II. INFRASTRUCTURE AND CHALLENGES OF IoT

A. *Infrastucture*

Hardware: IoT has extended its tentacles in terms of hardware across multi-protocol, multi standards readers, sensors, actuators and secure low-cost tags (silent tags). Smart sensors or bio-chemical sensors which are evolving may be embedded in the future IoT systems with tiny sensors with expandable long ranges or wireless power with the capacity of detecting animate objects more efficiently. A strive for energy efficient data processing and context adaptable systems for distributed registries, search and discovery systems which include discovery of sensors and

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.51

retrieval of sensor data on IoT systems are explored and adapted consistently.

Network Connectivity: Wirelessly networked sensors in IoT form a new Web as the data from these are collected constantly, analyzed and interpreted in a meaningful way. There are huge quantities of data which are collected using these sensors and the messaging volume could easily reach between 1.000 and 10.000 per person per day as they are constantly on [2].

Moreover IoT has made its presence in adhoc network formation, self-organizing wireless mesh networks, sensor RFID-based systems [5][8], Networked RFID-based systems by interfacing with hybrid systems and networks. They all need multi authentication to have a secure system and the communication between them should be regularly monitored as they will evolve into service based network and will start brokering data through market mechanism.

Smart systems on tags, with sensing and actuating capabilities (temperature, pressure, humidity, display, key-pads, actuators etc.) stores information and governing them is one of the challenges IoT systems face.

Architecture: In terms of IoT architecture extranet of things like partner to partner applications, basic interoperability and billions-of-things are the ones to be considered when designing such systems. SOA based IoT systems focus on the compositions of other services, single domain and single administrative entity and will be evolving into multi cross domain multi administrative entities and totally heterogeneous service infrastructures.

Software and Algorithms: Algorithms such as super algorithms which are installed in the new self-learning cars can predict the user's journey and map digital journey effectively. The security of these systems are very vital to clients' privacy and intrusion. IoT devices could pose security risk and difficulties if they are in sensitive areas where the information can be easily hacked into, resulting in opening the door to law suits.

There are known cases where the information held in the refrigerators have been taken over by hackers to install spambots. These super algorithms are used for tracing and context aware data distribution systems and is used to develop self-aware, self-management, self-configurable and self-healing systems. Bio-inspired algorithms based on game theory are one of the avenues IoT gaining momentum used in various networking and dynamic evaluation of systems like self-organization environments.

Cloud resource management's core functionality is to assign optimal resources for services on demand in paving way to self-aware or self-organized systems in dynamic environments [3][4]. Modelling and designing such systems involve complex algorithms and constant data mining, feedback from the sub systems. Self-aware systems built for energy efficient [7] "green cloud" operates by switching off and on the systems based on the usage is one of the IoT systems where continuous feedback or loopback of data is crucial part for day to day operations.

There are symmetric encryption public key algorithms and encryption is available for activating and deactivating tags as a means to protect sensitive objects/information. Object intelligence is also one of the areas which are evolving into IoT systems.

B. Compliance

Facebook, Google are using Drones which connects to cloud easily to track and are non-obtrusive. Such data collected should be subjected to data governance especially sensitive data which can be sold to companies for any purpose.

Many engineers and product developers are constantly looking for the development of industry standards or updates for these systems. The ISA[12] are at the forefront on the management of intelligent devices but an international standardization [9][10] for such devices for compliance, security protocols and effective data governance, storage of such data should be expanded and revised regularly as the technology spreads. Such standards should include communication within and outside cloud, data traceability, data creation and integrity. The IoT standardization should include M2M standardization, interoperability profiles and standards for cross interoperability with heterogeneous networks. The future of such standardization will be focusing on standards for automatic communication protocols.

C. Cloud Computing and IoT

Cloud computing is one of the dynamic environments available to all aspects of industries in the guise of software as a service, on demand services, infrastructure as services etc. The attractiveness of the cloud is its dynamic scalability and elasticity at multiple granularity features. There is an inherent understanding that resources in cloud computing are of type that can host and process data which involves processors, storage and virtualization. Even then, the sensors involved in the cloud aspects to provide enhanced capabilities related to reliability which are entirely dependent on feedback data will need information from these sensors which are in cohorts with Internet of things to scale resources dynamically.

Cloud technology offers vital support to IoT in terms of dealing with flexible amount of data originating from diversity of sensors and "smart things/objects" with the concept of scalability/elasticity to cope with dynamic data streams and mining [4].

In some systems the data collected from IoT systems are so vast that they have to be decentralized in the sense that the data should be distributed across the network and have efficient data indexing and mining techniques in place to retrieve them successfully. By decentralizing the storage of

data, the challenge of handling the big data generated by IoT can be handled efficiently.

II. QUALITY OF SERVICE IN IOT

QoS factors have been used in literature since the early hierarchical quality models [13] [14]. However, people tend to resist plans which evaluate many quality factors, due to limited resources or tight schedules. Based on previous research [13] the number of QoS factors should be kept between three and eight. Thus, in IoT QoS model only important and relevant factors will be considered.

Security: It is probable that the IoT could have challenges in the security aspects of the data transmitted between the devices and the mobility of the security mechanisms extending from firewall to authentication services [6]. For example a home IoT system is connected via a singular access point. The information between the systems are not secure and can be easily gathered. Smart Phone tracking data can be used for mapping digital footprints of a user which is easily stored in the stores servers. The apps downloaded in a user's smart phone device track users' locations, behavior patterns paving way for the misuse or selling of vital data to companies aimed at targeting customers for specific products. Existing privacy policies focus on data processing and virtualization and anonymization but a context centric security is needed according to the privacy needs based on automatic evaluation.

In the future it might include self-adaptive security mechanism and protocols and there may be standardization of IoT systems to reflect them. Moreover, companies like google who are using their predictable algorithms to store the user journey can track user's future progress and actions.

Performance: Performance of IoT depends on the scale of data existing in the system as it collects data from sensors, connected devices, cloud performance where it stores, network, signal strength and the frequency of the collection. It can only be fine tunes on device-by-device basis and the performance monitoring process needs to be amended to reflect the changes or the subsystem data. In terms of monitoring hardware IoT systems elements like the smart energy and power systems, wireless systems, smart racks, security sensors are to be included in the measurement and monitoring metrics. As they are dynamic an optimal strategy needs to be drawn for all the connected devices for the system to work efficiently.

Usability: is another important factor which is defined as the ability of a product/system to be used for the purpose chosen. It is a factor that is also considered important in many quality models. If a system isn't usable, then there is little point in its existence. To be useful, a system should have adequate documentation and support, and should have an intuitive easy-to-use interface.

Reliability: is defined as the probability of failure-free software operation for a specified period of time in a specified environment. Reliability is hard to achieve, because the complexity of software tends to be high. While any system with a high degree of complexity, including software, will be hard to reach a certain level of reliability, system developers tend to push complexity into the software layer, with the rapid growth of system size and ease of doing so by upgrading the software.

Robustness: is defined as the ability of a system to maintain its performance under undue pressure and changes. Not all systems are required to meet such factor, however in a distributed environment like Cloud, the possibility of predicting the number of users at any time can be a daunting task and therefore the need for the system to be able to perform when users increase and demand for services also increase.

Interoperability: is defined as the ability for a system to communicate and exchange information between each other and external systems of different structure.

Scalability: is the ability to increase a system without affecting the level of performance of a system and it is linked to performance which is an indication of how well a system, already assumed to be correct, works

Having identified the main QoS we proceed to create a hierarchal model of the QoS arranged on different levels, e.g. level A (LA) for high level factors, and subsequently the factors of the lower levels B (LB) and C (LC) (Figure 2) which show how QoS factors can be estimated.

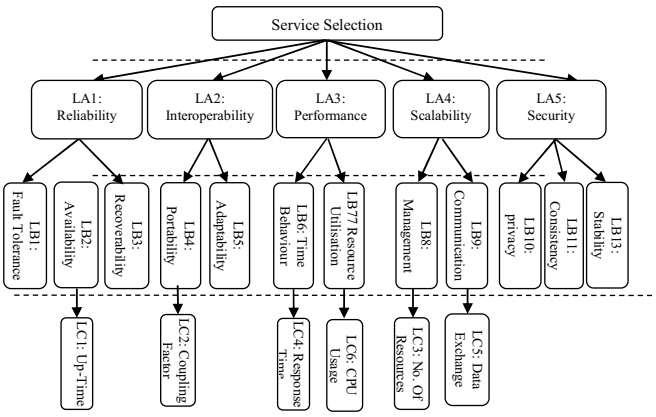


Figure 2 Multilayer QoS aspects

IV Conclusions and Future Work

The primary and secondary growth impact of IoT on various industries are tremendous and will continue to do so. Human and computer interaction systems and M2M communication systems do affect functional and non-functional requirements of a software product. Any quality model designed for such systems should take into account all these aspects and any metrics used for measuring will be a combination of metrics for individual quality factors selected to be crucial for any quality model. The challenges IoT faces and the standardization of IoT systems will be constantly evolving to keep up with the changes in both the technology and the areas it affects. There is a huge emphasis on big data and data mining techniques as the data gathered from these devices and systems are constantly on and monitored. They are mined to specific incidents or behaviour to trace the map or journey paving way to self-healing/awareness system.

Companies like Cisco [11] are researching into self-healing hardware which automatically corrects itself for known faults. In the future we can see a trend in all the faults or defects being stored and mapped to a specific fault of defect and there will be a work around solution for the system to be capable of performing all or nearly all the functionalities. The quality model designed for such systems should include external factors like signal strength and network connectivity in its non-functional requirements criteria as bandwidth spectrum needs to be measured and fine-tuned to exchange the massive amount of messages and information being transferred in and out of such systems. Finally, there should be constant focus on the security and compliance standard of such systems to safe guard the privacy information for data traceability.

References

[1] Shanzhi Chem, Hui Xu, Dake Liu, Bo Hu, Hucheng wang, 2014. A vision of IoT:Applications, Challenges, and opportunities with China Perspective, IEEE Internet of Things Journal Volume 1, No.4.

[2] Internet 3.0, 2010:The internet of thing © Analysys Mason Limited.

[3] Jay Kiruthika, Souhiel Khaddaj, 2014. Dynamic resource allocation:cloud computing, Proceedings of DCABES - IEEE conference publications in China.

[4] The future of Cloud Computing, Opportunities for Europe and Cloud Computing beyong 2010 Online. cordis.europa.eu/fp7/ict/ssai/events-20100126-cloud-computing_en.html.

[5] Li Da Xu; Wu He; Shancang Li, 2014. "Internet of Things in Industries: A Survey," Industrial Informatics, IEEE Transactions on , vol.10, no.4, pp.2233,2243.doi: 10.1109/TII.2014.2300753.

[6] Sylvain Kubler, Kary Främling, Andrea Buda.2015. A standardized approach to deal with firewall and mobility policies in the IoT, Pervasive and Mobile Computing, Volume 20, Pages 100-114, ISSN 1574-1192, http://dx.doi.org/10.1016/j.pmcj.2014.09.005.

[7] Keke Gai, Meikang Qiu, Hui Zhao, Lixin Tao, Ziliang Zong.2015. Dynamic energy-aware cloudlet-based mobile cloud computing model for green computing, Journal of Network and Computer Applications,ISSN 1084-8045, http://dx.doi.org/10.1016/j.jnca.2015.05.016.

[8] Taekyung Kim; Chenglong Shao; Wonjun Lee.2015. "Promptly pinpointing mobile RFID tags for large-scale Internet-of-Things," Big Data and Smart Computing (BigComp), 2015 International Conference on , vol., no., pp.118,123, 9-11 Feb. doi: 10.1109/35021BIGCOMP.2015.7072820.

[9] ISO/IEEE 11073:CEN Health informatics – Medical/health device communicatin standards

[10] Gomes, Y.F., Santos, D.F.S., Almeida, H.O., Perkusich, A.2015. "Integrating MQTT and ISO/IEEE 11073 for health information sharing in the Internet of Things," IEEE International Conference on Consumer Electronics (ICCE), pp.200,201.doi: 10.1109/ICCE.2015.7066380.

[11] Carlos Ramos, Paulo Novais, Céline Ehrwein Nihan, Juan M. Corchado Rodríguez.2014.Ambient Intelligence - Software and Applications.5th International Symposium on Ambient Intelligence. Advnaces in Intelligent Systems and computing Vol (291). ISBN: 978-3-319-07595-2

[12] ISA 108, 2012: Intelligent Device Management. https://www.isa.org/isa108/

[13] Khaddaj, Souheil, and Gerard Horgan. "The evaluation of software quality factors in very large information systems." Electronic Journal of Information Systems Evaluation 7.1, pp 43-48, 2004.

[14] G. Horgan, S. A. Khaddaj, "Use of an adaptable quality model approach in a production support environment" in 'Journal of Systems and Software', 82(4), Elsevier, April, pp. 730-738, 2009.

ANN Based High Spatial Resolution Remote Sensing Wetland Classification

KE Zun-You
School of Earth Sciences and Engineering, Hohai University, & Information Engineering Dept., Nanjing Institute of Mechatronic Technology, Nanjing, PRC
2265255075@qq.com

AN Ru
School of Earth Sciences and Engineering, Hohai University, Nanjing, PRC
anrunj@163.com

LI Xiang-Juan
College of Business Administration, Nanjing University of Traditional Chinese Medicine Nanjing, PRC
xjlee2002@126.com

***Abstract*—RS(Remote Sensing) image classification based on ANN(Artificial Neural Network) is carried out with high spatial resolution images of the wetland, which is the most important ecological environment element within the land components. Wetland dynamic change monitoring is often built upon its classification result concerned here. The typical high spatial resolution image of the wetland in Nanjing is used as a study case by ANN method in comparison with MLC(Maximum Likelihood Classification). Furthermore, the optimal number of ANN hidden neurons are simulated for enhance the classification effectivity. Totally, the results show classification method of ANN with optimal hidden neurons can effectively distinguish ground objects and improve the classification accuracy. The overall accuracy of the ANN classification is up to 93% and the Kappa coefficient is over 0.89.**

Keywords- Wetland Classification; Artificial Neural Network; Remote Sensing; High Spatial Resolution Image; Hidden Neuron Number

I. INTRODUCTION

As the nature's most species-rich ecological landscape, wetlands are the most important environment to human living. Wetlands have very high production capacity and the maintenance of ecological balance, such as flood and drought, freshwater preservation, climate regulation and protection of species diversity and other functions.

Remote sensing is an earth observation technology with a wide detection range, fast speed data acquisition for the study of surface changes providing a multi-platform, multi-spectral, multi-phase, a wide range of real-time information. Accuracy automatic remote sensing image classification method is the prerequisite and basis of other practical applications.

Artificial neural networks (ANN) are adaptive nonlinear dynamic system with a lot of simple neurons connected together. Its application has been extended to various engineering fields, with a strong self-learning, fault tolerance and the ability to complete computation, identification and control. Nonlinear mapping or adaptive dynamic system features of ANN can solve common Remote Sensing(RS) image processing difficulties, so it is quickly applied widely in the field of RS image classification.

II. REVIEW

RS image classification is performed by computing and analyzing spectral information and spatial information of ground features in RS images, selecting the features and using certain means to divide the feature space into non-overlapping sub-space. Then the individual pixel of the image classified to each sub-space. Some scholars have done studies of wetland classification and monitoring in a long sequence [1] , others researched classification using LiDAR texture and ANN[2] .

It's quite serious that different objects have approximate spectral information in pixel-based classification. The accuracy is to be improved [3,4]. ANN also used in sub-pixel classification[5]. With the development of ANN theory and high spatial resolution RS images, neural network technology is increasingly becoming an effective means of RS image classification processing. RS image classification based on neural network technology is applied effectively in aspects of land cover, crop classification and prediction of geological disasters, etc. [6,7].

Artificial neural networks are not by fractional arithmetic or logic to solve complex, but by adjusting the weights for the network connections of neurons, using non-algorithmic form of unstructured achieved.

ANN is not implemented with fractional-step algorithm or complex logic programs, but it is achieve by adjusting the connection weights of the neurons in the network and using non-algorithms, unstructured form.

ANN has been used to categorize a variety of RS data and its results are superior to traditional statistical methods. These successes can be attributed to neural network in nonlinear system of two major advantages [8]:

- The data is not required to distribute normally;
- Adaptive simulation of complex nonlinear models with a particular topology functions.

III. DATA AND METHODOLOGY

A. Region and Data

The study region is Nanjing Jiangxinzhou wetland park. This area shown in Figure 1 includes a number of beaches and water-based artificial and natural mixed wetlands.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.52

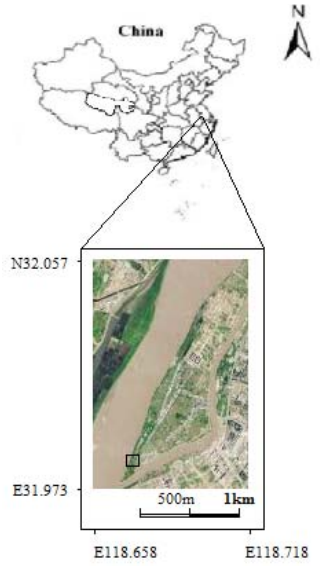


Figure 1. Study area location

The studied wetland system shown in Figure 2 contains a wealth of information, mainly water and wet beach information. The following high spatial resolution RS image data is processed and analyzed in this study.

Figure 2. Original study area remote sensing image

B. *Classification Methodology*

In BP(Back Propagation) algorithm ANN, there are some hidden layers between the input layer and output layer. Its neurons of adjacent levels are connected with specified weights and thresholds. Neurons in the same layer are not connected[9].

BP neural network used for remote sensing image classification consists of two processes:

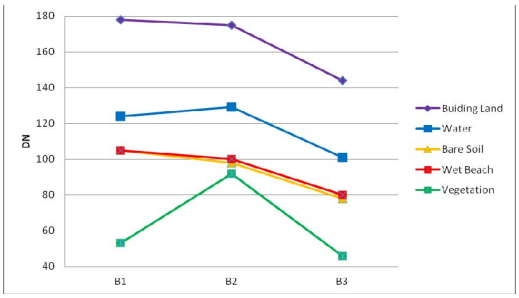
The neural network model construction and neural network simulation.

- Construction of neural network model is using Visually interpretation of known category attribute samples, training the neural network model, and establishing neural network model.
- Neural network simulation means each image pixel information is input into the trained neural network for processing pattern recognition and obtaining an output vector. The output vector is corresponding to the pre-determined category space. Then each pixel is determined to belong to specific subspace, and the purpose of automatic classification of remote sensing images is achieved.

Each category sample training pixels can be set through visual interpretation, and target parameters of each category is determined and into the network for training. Finally the simulation is done with wetland image data in the trained neural network, and each pixel category is determined in accordance with pre-determined each category range value to achieve the purpose of classification.

1) Classification system

Wetlands contain natural or artificial, permanent or temporary wetlands, peat land or water areas, still or flowing water. Wetland classification is the basis for the study of wetland ecosystems. The paper is based on the concept of wetland, referring to "Convention on Wetlands" (1990) as well as domestic and foreign study used classification system and specific applications of the wetland ecological research system, taking into account the RS data sources interpretation[10]. The research land use / cover classification system is established: building land, vegetation, wet beach, water and bare soil, where the wet beach and water belong to the wetland type. Their DN



value are shown as Figure 3.

Figure 3. Typical ground objects' spectral information DNs

2) ANN training samples

The RS classification selected sample should have a representative and typical geography. Moreover, the spectrum, the number of training samples in the study area should be based on the region size of each category: more water and less building lands and wet beach samples.

Selection of samples are through visual interpretation of RS image.

Figure 4. Sample selection in the study area

3) Wetland classification experiments

Neural network classification uses neural network toolbox in Matlab 8.0.

Since the common BP algorithm has defects of slow convergence and easy to fall into local minimum, we use additional momentum and adaptive learning rate improved BP algorithm in a 3-layer BP algorithm ANN.

The neurons number of hidden layer in the neural network model relates to network abilities of learning and generalization in simulation. Empirical algorithms of hidden layer neuron number [8,11] and adaptive optimization algorithm [12,13] are put forward.

Here neural networks hidden number refers to the empirical formula as the following[8]:

$$S=R+SL+a \quad (1)$$

Wherein,

Parameter S is the number of hidden layer nodes.

Parameter R is the number of input nodes.

Parameter SL is the number of output layer nodes.

Parameter a is an empirical constant in the range of 0 to 10, obtained by trial.

Generally, the more simple the ANN structure is, the better generalization ability it has. In order to prevent over-learning neural network, the size of the ANN should be minimized based on the simulation and prediction.

The target ANN has three layers: input layer with 3 neurons, hidden layer with 4-20 neurons and output layer with 1 neuron.

In Matlab ANN maximum training is set to 1000 times, training error is 0.1, the learning rate is 0.01. The ANN training target parameter, range of values of each category for neural network output layer and pseudo-color are shown in Table I.

TABLE I. ANN CLASSIFICATION CATEGORIES TARGET, RANGE OF VALUE AND PSEUDO-COLOR

Category	Building Land	Water area	Bare Soil	Wet Beach	Vegetation
Training target	1	2	3	4	5
Range of value	<1.5	(1.5, 2.5)	(2.5, 3.5)	(3.5, 4.5)	>4.5
Pseudo-color	Purple	Blue	Yellow	Red	Green

4) Number of hidden layer neurons

The ANN based classification pseudo-color images can be drawn out with the parameter of different hidden layer neurons (S = 4,5,6, ... 20) .

5) Comparison and accuracy analysis

Maximum Likelihood Classification (MLC) is also carried out in ENVI.

Single hidden layer BP algorithm neural network typical Confusion Matrix of classification is to be analyzed. Here the classification result of S=5 is taken into account for demonstrating the confusion and accuracy, and compared with MLC implementation result.

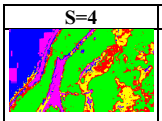
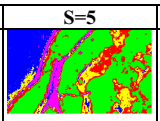
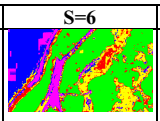
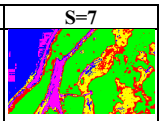
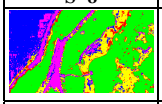
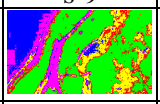
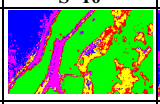
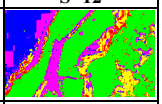
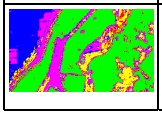
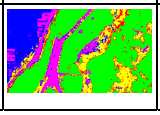
IV. RESULT AND DISSCUSION

RS image spectral information classification based on ANN with different parameter of S and MLC are processed, and compared as to Nanjing Jiangxinzhou wetlands high spatial resolution RS image.

1) Classification results of ANN with different hidden layer neurons

The pseudo-color images of ANN based classification after processed in Matlab with different number of hidden layer neurons (S = 4,5,6 ... 20) in 3-layer forward neural network are shown in Table II:

TABLE II. CORRESPONDING ANN CLASSIFICATION PSEUDO-COLOR IMAGE OF DIFFERENT HIDDEN LAYER NEURONS NUMBER

S=4	S=5	S=6	S=7
			
S=8	**S=9**	**S=10**	**S=12**
			
S=15	**S=20**		
			

- From the ANN classification results of different number of neurons in the hidden layer, generally the main area classification basically the same. However, when S>10, a large number of wet beach (red) data loss.
- From the view of the smaller local area of the study area, Yangtze water (blue) in most classification result image is wrongly divided into building land (purple) except while S = 5.
- Other small areas, such as building land surrounded by vegetation and green island surrounded by water, etc., are classified poorly while S = 4,6,7.
  The shadows of big trees are often classified wrongly to wet beach (red), which could be avoided mostly if original image preprocessed taking into account the solar altitude.
- As a whole, while S = 5, the classification correctness is higher than others, since the classifications balance better between the large area and the details of the image.

2) Comparison and accuracy analysis

Here the ANN classification result of S=5 is taken into account. Single hidden layer BP algorithm ANN Confusion Matrix of classification is shown in Table III.

TABLE III. BP ALGORITHM NEURAL NETWORK TYPICAL CONFUSION MATRIX BY GROUND TRUTH (PERCENT)

Class	Wet Beach	Vegetation	Bare Soil	Water	Building Land
Wet Beach	12.58	0	2.05	0	0.42
Vegetation	0.12	100	0	0	0
Bare Soil	86.43	0	97.95	0	0.08
Water	0.87	0	0	99.98	7.49
Building Land	0	0	0	0.02	92.01
Total	100	100	100	100	100

Single hidden layer BP algorithm neural network classification accuracy is shown in Table IV.

TABLE IV. ANN CLASSIFICATION ACCURACY

Class	Prod.Acc.	User Acc.
Wet Beach	12.58	75.94
Vegetation	100	99.96
Bare Soil	97.95	65.02
Water	99.98	98.43
Building Land	92.01	99.91
Overall Accuracy = 93.23% Kappa Coefficient= 0.8985		

The overall accuracy of the ANN classification is up to 93% and the Kappa coefficient is over 0.89.

The classification pseudo-color image of MLC after processed in ENVI is shown in Figure 5:

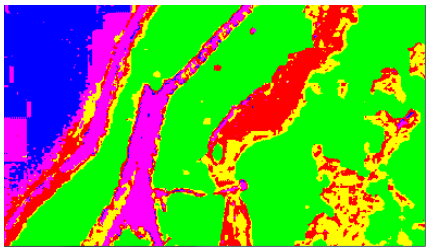


Figure 5. Pseudo-color image of MLC

The comparison Confusion Matrix of ANN and MLC classification is shown in Table V, and the accuracy of comparing the two classifications is shown in Table VI.

TABLE V. THE COMPARISON CONFUSION MATRIX OF ANN AND MLC CLASSIFICATION

Class	Wet Beach	Veg.	Bare Soil	Water	Building Land	Tot.
Wet Beach	46.78	1.52	21.69	0.71	0	13.34
Vegetation	16.54	97.29	0	0.32	0	50.84
Bare Soil	34.33	1.09	77.75	1.01	2.24	12.33
Water	0	0	0	77.21	0	11.71
Building Land	2.35	0.1	0.56	20.75	97.76	11.79
Total	100	100	100	100	100	100

TABLE VI. ACCURACY OF COMPARING THE TWO CLASSIFICATIONS

Class	Prod. Acc.	User Acc.
Wet Beach	46.78	87.69
Vegetation	97.29	91.77
Bare Soil	77.75	23.42
Water	77.21	100
Building Land	97.76	67.74
Overall Accuracy = 80.9286% Kappa Coefficient = 0.7240		

As a whole the classification result of ANN consists with the one of MLC. Respectively the concordance rate of building land and vegetation is even over 97%.But in MLC, over 20% of the water is mistaken for building land, and more than 20% of bare soil is mistaken for wet beach. In addition, wet beaches and bare soil vary more obviously between ANN and MLC, especially concordance rate of wet beach being less than 47%.

V. ARGUMENTATION AND CONCLUSION

Because of the complexity of wetland ecosystem, the same ground objects would be different in spectral information, for instance, waters containing different quantity of sands differ in spectral characteristics.

Wetland of wet beach includes complex ground feature information, different objects with quite the same spectra information are very common. So their classification based upon feature spectral information always has obvious mistakes.

From the above simulation, as to BP algorithm ANN classification, while hidden layer neuron number S = 5, the classification has better correctness of a large area and better balance in local details. As a whole the classification result of ANN consists with the one of MLC.

ACKNOWLEDGMENT

The authors would like to thank to the National Natural Science Foundation (NNSF) project 41271361.

REFERENCES

[1] Yang Ren-Min, An Ru， Wang Hui-Lin， Chen Zhi-Xia， Quaye-Ballard Jonathan， “Monitoring Wetland Changes on the Source of the Three Rivers From 1990 to 2009, Qinghai, China”， IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing， vol.6, no. 4, pp.1817-1824, Jul. 2013.

[2] QIAO Jigang,LIU Xiaoping, ZHANG Yihan, “Land cover classifi cation using LiDAR height texture and ANNs, ” Journal of Remote Sensing,vol.15, no.3, pp.539-553, Mar. 2011.

[3] JIANG Hui1, ZHOU Wen-bin, LIU Yao, “Research and Application of the Poyang Lake WetLand Classification Using Remote Sensing,” Remote Sensing Technology and Application, vol.23, no.6, pp.648-652, Dec.2008.

[4] LI Xiao-dong1, GUO Zhong-yang, ZHU Yan-ling, DAI Xiao-yan, “Artificial neural network classification of wetland integrating GIS Data,” Journal of East China Normal University(Natural Science), no.4, pp.26-33, Jul. 2010.

[5] Karkee Manoj, Steward Brian L., Tang Lie, Aziz Sarnsuzana A., “Quantifying sub-pixel signature of paddy rice field using an artificial neural network,” COMPUTERS AND ELECTRONICS IN AGRICULTURE, no.65, pp.65-76, Jul.2008

[6] BERBEROGLU S, CURRAN P J, LLOYD C D., “Texture classification of Mediterranean land cover,” International Journal of Applied Earth Observation and Geoinfomation, no.9, pp.322-324, Nov.2007.

[7] MAHESH PAL, PAUL M M, “An assessment of the effective of decision tree methods for land cover classification,” Remote Sensing of Environment, vol.86, pp. 554-565, Aug. 2003.

[8] Zhou Kaili, Kang Yaohong, Neutral Network Model and Matlab Simulation Program Design, CN: TsingHua University Press, 2005.

[9] Martin T. Hagan et. al., Neutral Network Design, CN： China Machine Press, 2002.

[10] WHIGHAM D.Wetlands of the United World. Dordrecht:Kluwer Academic Publishers,1993.

[11] GaoPengyi， Study on The Optimization of Backpropagation Neural Network Classifier，CN:Huazhong University of Science and Technology, 2012.

[12] Wang Liwei, Determination of Number of ANNs’ Hidden Layer Neurons，CN:Chongqing University, 2012.

[13] ZHAO Linming,WEI Dehua, “The method to choose the optimal number of hide nodes of artificail nueral networks,” Journal of North China Institute of Water Conservancy and Hydroelectric Power, vol.20, no.4, pp.44-48, Dec. 1999

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A new way of combining RDP and Web Technology for mobile virtual application

Yousong Zhang, Wei Wei, Pengdong Gao, Yongquan Lu, QuanQi
Communication University of China
Beijing, China
e-mail: wwwzys@163.com

***Abstract*—As the important part of the virtual application system, the remote control software plays an important role in the interaction between system and user. With the rise of mobile Internet, the virtual desktop software which runs in PC has been very difficult to adapt to the demand for mobile communication. This paper mainly introduces mobile virtual application, which based on the Remote Desktop Protocol and learned from the excellent open source project, is suitable for cultural and creative platform. The biggest advantage of software is the combination of Web Technology, simplifying the user tedious connection configuration parameters, and optimize the connection speed, improved the user operability.**

Keywords-component; RDP; mobile virtual application; URI Scheme;

I. INTRODUCTION

The remote connection software generally refers to the software that can remotely connect to a computer, used to monitor or control the computer. Especially for a computer, opened the remote desktop connection service, it can control of the computer through the remote desktop on the other side of the network, operate the computer and install software on top, run the program real time. The mobile terminal of the remote connection software is equipped with such software on mobile devices, remote connection software of mobile terminal with the characteristics of mobile software, easy, convenient and breaking the restriction of bulky PC, basically can be used in a variety of convenient network environment, including Wi-Fi, mobile network environment.

The platform of creation service for the cultural and creative can provide all kinds of software resources for the creators, which can make creator, create works online. The creation and exhibition platform online is based on Web, and creators of the authoring software is provided by the platform. Platform virtualization technology, cloud technology [1-5] provides virtual software services, through the mobile terminal remote connection software, allowing creators to smooth using cloud virtual terminal software services. Due to run the software in the mobile terminal, greatly facilitate the creator, can create works at anytime and anywhere timely, so the web and mobile app combined to become the key and difficulty of the system.

Now the popular remote connection the software, mostly runs in the PC-side, really popular mobile terminal software is rare. And this software has multifarious connection parameters for setting, for the average non-computer professional people in the field of ordinary user. It is difficult to use the software. Multifarious parameters become a big obstacle for using remote connection software. Especially in the online creative platform, if allowing the user to enter the complex parameters will lose the enthusiasm of users for using it. Therefore simplify operation, reduce the connection parameters, can improve the user's using enthusiasm, more conducive to software, as well as the corresponding platform promotion. In addition, The Remote Desktop Protocol speeds up the transmission of images ultimately through the establishment of virtual channel, establishing channel takes a certain amount of time, however, the corresponding processing, used here to accelerate channel set up time, can make a connection time is shorter, so as to avoid the user for a long time of waiting, to achieve a better user experience.

II. RELATED WORK

A. Virtual application

Virtual application or application virtualization, realizes “software as a service” (SaaS) through virtualization technology. SaaS is the most mature, famous and popular cloud computing type, it is the cloud service provider provides the application software services through the Internet, and the users get the application service from the provider by pay-on-demand [6].

B. Mobile Virtual application

Mobile Virtual application is based on Virtual application, which can connect remote virtual system by the mobile network. It makes virtual system use scope to further expand.

C. RDP protocol

Remote Desktop Protocol (RDP) is a proprietary protocol developed by Microsoft, which provides a user with a graphical interface to connect to another computer over a network connection. The user employs RDP client software for this purpose, while the other computer must run RDP server software.

Based on the ITU-T T.128 application sharing protocol (during draft also known as "T.share") from the T.120 recommendation series, the first version of RDP (named version 4.0) was introduced by Microsoft with "Terminal Services", as a part of their product Windows NT 4.0 Server, Terminal Server Edition. The Terminal Services Edition of NT 4.0 relied on Citrix's MultiWin technology, previously provided as a part of Citrix WinFrame atop Windows NT 3.51, in order to support multiple users and login sessions simultaneously. Microsoft required Citrix to license their

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.53

MultiWin technology to Microsoft in order to be allowed to continue offering their own terminal-services product, then named Citrix MetaFrame, atop Windows NT 4.0. The Citrix-provided DLLs included in Windows NT 4.0 Terminal Services Edition still carry a Citrix copyright rather than a Microsoft copyright. Later versions of Windows integrated the necessary support directly. The T.128 application sharing technology was acquired by Microsoft from UK software developer Data Connection Limited.[7]

D. RDP working mechanism

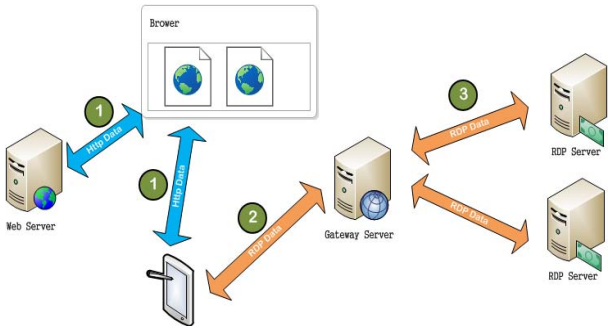


Figure 1 RDP Network Structure

1) RDP Network Structure

(1) Users browse web page that contains URI.

(2) Users only click URI of web page and local app about RDP will run and connect the gateway.

(3) The gateway will communicate with RDP server and transfer RDP data to the users.

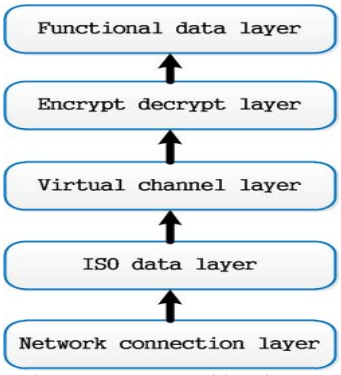


Figure 2 RDP working layers

2) RDP working layers

a) Network connection layer

RDP protocol based on TCP/IP protocol, due to the amount of data transferred is large, so in the underlying agreement first define a network connection layer. It defines a complete package of RDP data in logic, in order to avoid that because of the length of the network packet is too long, the data is lost.

b) ISO data layer

ISO data layer is on the network layer, which represents the RDP normal connection communication of data.

c) Virtual channel layer

Virtual channel layer is on the ISO data layer, RDP protocol defines a virtual channel layer, to break up the data of different virtual channel, to speed up the client processing, save for the network interface time.

d) Encrypt decrypt layer

Encrypt decrypt layer is on the virtual channel layer, RDP defines a data encryption to decrypt layer. This layer is used to treat all data encryption and decryption processing function.

e) Function data layer

The functional data layer is on the encryption decryption layer, The encryption decryption layer is the functional data layer, Transform for image information, the local resources, voice data, print data, and so on all the function of data information in this layer for processing. In addition, according to different types of data, these data have different levels of segmentation; their internal hierarchy will be done in each function module in detail.

E. FreeRDP

FreeRDP is a free open-source of the Remote Desktop Protocol (RDP), it also has high extensibility and re-configurability, it currently supports Windows and Linux, and can be ported to other OS's including OS X, IOS and Android. [8]

The software designed and developed based on the framework.

III. IMPLEMENTATION

A. URI Scheme

A URI scheme is the top level of the uniform resource identifier (URI) naming structure in computer networking. All URIs and absolute URI references are formed with a scheme name, followed by a colon character (":"), and the remainder of the URI called the scheme-specific part. The syntax and semantics of the scheme-specific part are left largely to the specifications governing individual schemes, subject to certain constraints such as reserved characters and how to "escape" them.

URI schemes are frequently and incorrectly referred to as "protocols", or specifically as URI protocols or URL protocols, since most were originally designed to be used with a particular protocol, and often have the same name. The http scheme, for instance, is generally used for interacting with web resources using HyperText Transfer Protocol. Today, URI with that scheme is also used for other purposes, such as RDF resource identifiers and XML namespace, which are not related to the protocol. Furthermore, some URI schemes are not associated with any specific protocol and many others do not use the name of a protocol as their prefix. [9]

URI schemes should be registered with IANA, although non-registered schemes are used in practice. RFC 4395 describes the procedures for registering new URI schemes. [10]

B. Parameter design

In order to solve the complex connection parameters, the user from multifarious connection parameters setting, URI

scheme must pass in enough parameter. The following design some necessary parameters, and set aside a part of the parameters as expansion.

1) hostname
The address of the server is connected

2) port
The port of the server is connected

3) username
The user name of the server is connected

4) password
The password of the user

5) domain
Domain address of the server is connected

6) enable_tsg_settings
Whether to enable gateway settings

7) tsg_hostname
Ip address or name of gateway server

8) tsg_port
Port of gateway server

9) tsg_username
The user name of gateway server

10) tsg_password
Password of gateway server

11) tsg_domain
The domain of gateway server

12) remote_program
Identifier of the application be opened

13) key
Ensure the authenticity of the parameters

C. URI generation

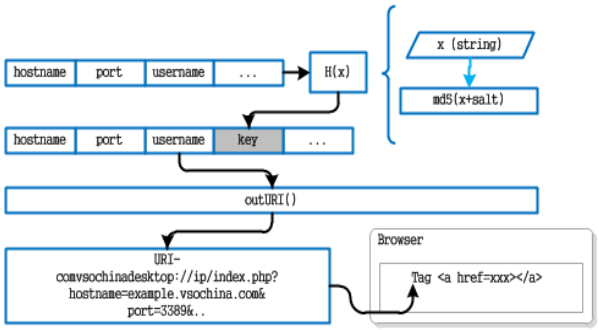


Figure 3 URI generation

Figure 3 illustrates the method for generating a URI. The main function of H(x) is to hash function operation, and mainly process the parameter. The key will be generated by H(x), and x is regular splicing of parameters. The output of H(x) is key field. The key as a field appended to the URI, thus generating the can verify the authenticity of the URI. H (x) is a hash function, can contain the md5 algorithm, or the other of the hash algorithm. The salt as a character string is used to enhance security can dynamically change over time.

An example of a URI:
comvsochinadesktop://desktop.vsochina.com/index.php?
hostname=example.vsochina.com&
port=3389&
username=30870310&
password= password &
domain=vsochina.com&
enable_tsg_settings=YES&
tsg_hostname=rdgwexample.vsochina.com&
tsg_port=443&
tsg_username=30870310&
tsg_password=password&
tsg_domain=vsochina.com&
remote_program=||wordpad&
key=asdhiuwdpqwpdacsd1adk22sjak

D. Parameters verification

To prevent malicious calls, and verify the authenticity of information, URI parameters passed using the calibration procedure. Through the check mechanism, can effectively identify the authenticity of user information connection information, to prevent illegal access. The check mechanism is shown in the figure below.

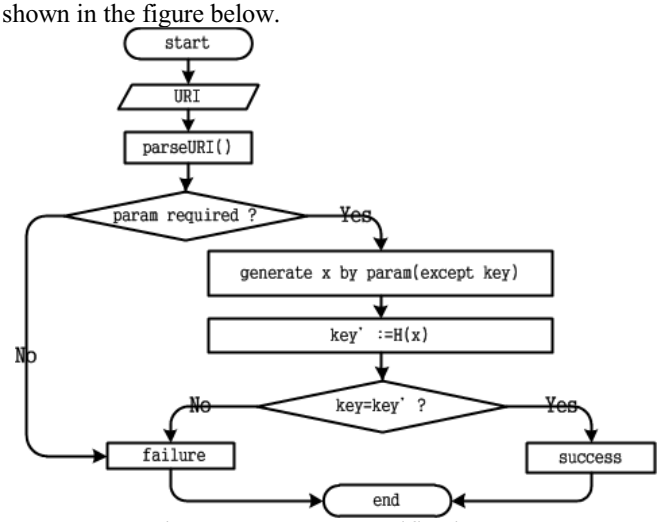


Figure 4 Parameters verification

ParseURI(), a function, aims to parse the URI for each parameter, the integrity of the first determines parameters. If the parameter is not integrated or not correct, program will return failure. If the parameter is integrated, program will be into the next phase of processing. The various parameters (in addition to the key field) will combine into a string x, the key' will be generated by function H(x), which is hash function. After key' generating, the following will judge key' and key is same or not. So far, the verification process is completed.

E. Expectations and the actual effect

Figure 4 Expected operation

Figure 4 shows the way of using, when users click the URI from any web browser, automatic call local app and

open the remote virtual application immediately, thus this eliminates the trouble of setting up a large number of parameters, compared to conventional remote software.

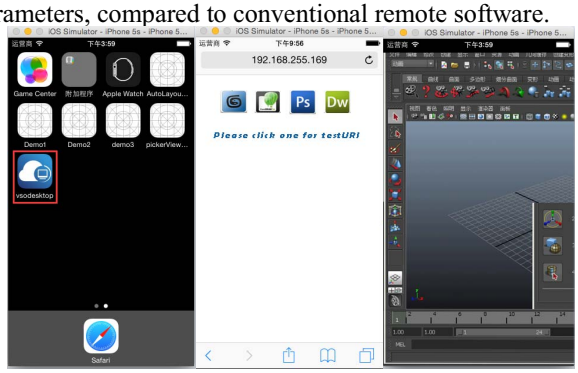


Figure 5 Actual operation

Figure 5 shows the IOS 5 Simulator running effect, through the Safari browser open platform to test, click on the corresponding URI, call native remote connection software, can automatically avoid complex settings.

IV. PERFORMANCE COMPARISON

There are some results of performance about remote software.

Implements	Connection delay(second)	Start delay(second)	UE score (avg)
This App	0.22	1.12	85
HTML5	0.35	1.43	64

Data in a table illustrates the app can spend less delay than HTML5, and the app can get better user experience score.

V. CONCLUSION AND FUTURE WORK

In this paper, we hope a solution to reduce the response latency and improve ease of operation. As can be seen from the above data, this application basically achieved the expected results. But the solution also has some deficiencies that need improvements in further work:

- Optimization of the connection parameters further
- Optimization algorithm reduces calibration check time
- Optimization other mobile platform

With the above approach, it can solve the binding problem between web platforms and applications about remote mobile operation system, and is ideal solution for cultural and creative platform.

ACKNOWLEDGMENT

The authors acknowledge the financial supports by the National Key Technology Support Program (2012BAH17B03) and the Program Project of CUC (XNG1138, YXJS2012319, YXJS2012206, BY2012230, BE2013054 and JSWHCY-2013-(98)).

REFERENCES

[1] R. Nathuji, A. Kansal, and A. Ghaffarkhah, "Q-clouds: managing performance interference effects for QoS-aware clouds," Proc. 5th European conference on Computer systems, ACM Press, Apr. 2010, pp. 237-250.

[2] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, et.al, "An analysis of performance interference effects in virtual environments," Proc.IEEE Symp. In Performance Analysis of Systems &Software(ISPASS'07), IEEE Press, Apr. 2007, pp. 200–209.

[3] O. Tickoo, R. Iyer, R. Illikkal, and D. Newell, "Modeling virtual machine performance: challenges and approaches," ACM SIGMETRICS Performance Evaluation Review, vol. 37, Dec.2009, pp. 55-60.

[4] Chen K, Zheng W M. Cloud computing: system instances and current research[J]. Journal of Software, 2009, 20(5): 1337-1348..

[5] L.Wang, R.Ranjan, J. Chen et al. Cloud computing: Methodology, systems and applications[M]. Boca Raton: CRC Press, 2012.

[6] Yongquan Lu, Pengdong Gao, Chu Qiu, Jintao Wang,"A service cloud platform based on high-performance computing," China Science and Technology Achievements, vol. 12, no. 10, pp. 520-531, 2011.

[7] https://en.wikipedia.org/wiki/Remote_Desktop_Protocol#cite_note-2

[8] FreeRDP Open Source Project https://github.com/FreeRDP/FreeRDP

[9] URI scheme https://en.wikipedia.org/wiki/URI_scheme

[10] RFC 4395 https://tools.ietf.org/html/rfc4395

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Kinematics of 3-UPU Parallel Leg Mechanism Used for a Quadruped Walking Robot

Qifang Gu

Dept. of Electronic Information Engineering, WuXi City College of Vocational Technology,Jiangsu,WuXi,214153

***Abstract*—Compared with series mechanism, the parallel mechanism has many advantages. Using the parallel mechanism as the basic leg mechanism of a walking robot, not only the payload-weight ratio can be improved, but also the robot walking stability and security performance can be enhanced. This paper proposes the application of a typical 3-UPU parallel mechanism in a quadruped walking robot. The inverse and forward position solutions of 3-UPU parallel mechanism as standing leg are obtained. Based on the first order kinematics influence coefficient matrix, the corresponding global velocity performance index of the 3-UPU parallel mechanism is analyzed. This research provides a theoretical basis of further investigation for the quadruped walking robot with parallel leg mechanism.**

Keywords-walking robot; parallel leg mechanism; position analysis; influence coefficient matrix; performance analysis

1. INTRODUCTION

In recent years, different parallel mechanisms with fewer than 6-DOF called limited-DOF mechanisms which maintain the inherent advantages of parallel mechanisms and possess several other advantages such as reduction of the total cost of the device, are attracting attentions of various researchers [1].At present, the lower-mobility parallel mechanisms used in research and application mostly are 3 DOF translational mechanisms and 3 DOF spherical mechanisms which are based on the evolution of DELTA mechanism [2]. In the past two decades, the domestic and foreign scholars have proposed a large number of the new space lower-mobility parallel mechanisms which have good application value [3-9].This paper introduced a three-DOF translational 3-UPU parallel mechanism, proposed the application of a typical 3-UPU parallel mechanism in a parallel walking robot, carried out position analysis, velocity analysis, and the corresponding global velocity performance index analysis based on the first order kinematics influence coefficient matrix. This research provided a theoretical basis of further investigation for the quadruped walking robot with parallel leg mechanism.

2. THE APPLICATION OF 3-UPU PARALLEL MECHANISM IN A QUADRUPED WALKING ROBOT

Tsai [10] proposed a 3-UPU parallel mechanism which has 3 DOF and can achieve three-dimensional translation. The 3-UPU parallel mechanism could position a platform in a certain three-dimensional space, and make the platform parallel with a specific plane. Paper [11] first applied the 3-UPU parallel mechanism to the quadruped walking robot used for the elderly and the disabled as shown in Fig. 1.

Fig. 1.The quadruped walking robot

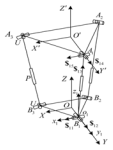


Fig. 2.The 3-UPU mechanism

When varying the orientation of joints, the 3-UPU parallel mechanism may have various DOF and kinematics characteristics. Fig. 2 shows the diagram of the 3-UPU parallel mechanism. The lower platform is defined as the fixed platform, and the upper platform is defined as the move platform. The branch coordinate system $o_1x_1y_1z_1$ on branch 1 is fixed at B_1. $\$_i(i=1,2,...,5)$ are the five spirals of each branch. From the assembly features of the 3-UPU parallel mechanism, the following relation can be written: $\$_1 // \$_5, \$_2 // \$_4$.

According to the requirements of DOF during quadruped walking, the leg mechanism should achieve planned gait with the movement of the body and can adjust the stability of the robot by its movement. In general, quadruped walking robot should achieve a three-dimensional movement at least, so the leg mechanism should meet the requirements of DOF during quadruped walking. The 3-UPU parallel mechanism shown in Fig.2 can realize three-dimensional translation [10]. The quadruped walking robot consists of one body and four identical parallel leg mechanisms, and the upper platform of each parallel leg mechanism is fixed with the body. When the lower platform of parallel leg mechanism contacts with the ground, the parallel leg mechanism is defined as standing leg; when the lower platform does not contact with the ground, it is defined as swing leg. In the walking process, the robot realizes the overall movement by lifting and falling of the four leg mechanisms in turn according to a certain gait order.

Compared with the series mechanism, using the parallel mechanism as the basic leg mechanism of a walking robot, it can enhance the walking stability and security performance, improve the payload-weight ratio, save the battery energy and prolong the walking time.

3. THE POSITION ANALYSIS OF 3-UPU PARALLEL LEG MECHANISM

A The position analysis of the standing leg

The position analysis of mechanism solves the relationship of the position between the input and output. The inverse position problem of the mechanism is to obtain each link length when the position and orientation of the moving platform in space are given. Conversely, it is the forward position problem of the mechanism.

If the 3-UPU acts as standing leg, the lower platform contacting with the ground is regarded as the fixed platform, and the upper platform moving with the body is regarded as the moving platform. As shown in Fig. 2, we establish a moving coordinate system $O'X'Y'Z'$ at the center of the moving platform. The joint centers of the upper platform connected with each branch are marked as $A_i'(i=1,2,3)$ in the moving coordinate system, and the joint centers of the lower platform connected with each branch are marked

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.54

with $B_i(i=1,2,3)$ in the fixed coordinate system. According to the actual designed size of the robot, set the circumcircle radius of the lower platform of the 3-UPU parallel mechanism as r and the circumcircle radius of upper platform as R. By the method of coordinate transformation, any vector A_i' in the moving coordinate system can be expressed in the fixed coordinate system as follows:

$$\boldsymbol{A}_i = [T]\boldsymbol{A}_i' \tag{1}$$

For the 3-UPU parallel mechanism, the movement of moving platform is the three-dimensional translation. The transformation matrix between upper and lower platforms is:

$$[T] = \begin{bmatrix} 1 & 0 & 0 & P_X \\ 0 & 1 & 0 & P_Y \\ 0 & 0 & 1 & P_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Using the geometrical relationship of regular triangle, all vertices of the triangle in the moving platform in the moving coordinate system can be expressed as follows:

$$\begin{aligned} A_1' &= [R \quad 0 \quad 0]^T \\ A_2' &= \left[-R/2 \quad \sqrt{3}R/2 \quad 0\right]^T \\ A_3' &= \left[-R/2 \quad -\sqrt{3}R/2 \quad 0\right]^T \end{aligned} \tag{3}$$

All vertices of the triangle in the fixed platform can be expressed as follows:

$$\begin{aligned} B_1' &= [r \quad 0 \quad 0]^T \\ B_2' &= \left[-r/2 \quad \sqrt{3}r/2 \quad 0\right]^T \\ B_3' &= \left[-r/2 \quad -\sqrt{3}r/2 \quad 0\right]^T \end{aligned} \tag{4}$$

According to equation (1), the coordinate value of hinge points in the moving platform can be obtained in the fixed coordinate system $OXYZ$. The length vector of driving link in the fixed coordinate system can be expressed as:

$$\boldsymbol{L}_i = \boldsymbol{A}_i - \boldsymbol{B}_i \quad (i=1,2,3) \tag{5}$$

The equation of inverse position solution can be obtained as follows:

$$L_i = \sqrt{L_{ix}^2 + L_{iy}^2 + L_{iz}^2} \qquad (i=1,2,3) \tag{6}$$

From the above equation, we have:

$$\begin{aligned} L_1 &= \sqrt{\left(R + P_X - r\right)^2 + P_Y^2 + P_Z^2} \\ L_2 &= \sqrt{\left(-\tfrac{1}{2}R + P_X + \tfrac{1}{2}r\right)^2 + \left(\tfrac{\sqrt{3}}{2}R + P_Y - \tfrac{\sqrt{3}}{2}r\right)^2 + P_Z^2} \\ L_3 &= \sqrt{\left(-\tfrac{1}{2}R + P_X + \tfrac{1}{2}r\right)^2 + \left(-\tfrac{\sqrt{3}}{2}R + P_Y + \tfrac{\sqrt{3}}{2}r\right)^2 + P_Z^2} \end{aligned} \tag{7}$$

Equation (7) is three independent explicit equations. When the fundamental dimensions of the mechanism, the location and posture of upper platform are given, we can make use of the upper equations to obtain the changes of the branch length of the three branches.If L_1, L_2, L_3 are given, the solution of P_X, P_Y, P_Z can be obtained. From equation (7), we can work out the relation equation between the origin position of the leg mechanism in the moving coordinate system and the input of each branch:

$$\begin{aligned} P_X &= \frac{1}{6(R-r)}\left(2L_1^2 - L_2^2 - L_3^2\right) \\ P_Y &= \frac{\sqrt{3}}{6(R-r)}\left(L_2^2 - L_3^2\right) \\ P_Z &= \pm\frac{1}{3(R-r)} \\ &\sqrt{L_1^2L_2^2 + L_1^2L_3^2 + L_2^2L_3^2 - \left(L_1^4 + L_2^4 + L_3^4\right) - 3(R-r)^2\left[3(R-r)^2 + \left(L_1^2 + L_2^2 + L_3^2\right)\right]} \end{aligned} \tag{8}$$

From equation (8), when increasing the input of each branch, we can obtain two groups of the origin position solutions in the coordinate system of moving platform. The value of P_X and P_Y are equal, and the value of P_Z is on the contrary. For 3-UPU parallel mechanism, the universal joints are arranged in the same direction at two ends of sliding joints of each branch, and the structure is symmetrical. When the 3-UPU parallel mechanism acts as standing leg, the lower platform is fixed, and the upper platform is moving. When it acts as swinging leg, the upper platform is fixed, and the lower platform is moving. Under the two cases, the values of P_Z are contrary.

If the 3-UPU acts as swing leg, the upper platform of the 3-UPU is relatively fixed with the body, and the lower platform acts as moving platform. The kinematics positive and inverse solutions of swing leg are similar to the process of solving the problem of standing leg.

B Position simulation of leg mechanism

When the 3-UPU acts as standing legs, set $R = 162mm$, $r = 58mm$, and the body move from the initial location (0,0,400) to the location (200,0,400), according to the equation (7), we can get the changing curve of the link length of all the branches in the process as shown in Fig. 3:

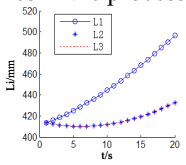


Fig. 3.The length changing curve of standing leg

It can be seen from Fig. 3 that when the 3-UPU mechanism acts as the leg mechanism of walking robot, the length of three branches changes smoothly. And variable quantity of branch 2 and branch 3 is always equal, and accord with actual movement.

4. THE VELOCITY ANALYSIS OF 3-UPU PARALLEL LEG MECHANISM

According to imaginary mechanism method, we add a nominal rotation pair $\$_6$ at the end of the branch. According to kinematic screw of various branches, we can obtain the first order influence coefficient matrix of all branches:

$$\left[G_\phi^H\right]_n = \begin{bmatrix} \left[G_\phi^h\right]_n \\ \left[G_\phi^P\right]_n \end{bmatrix} \tag{9}$$

where, $\left[G_\phi^h\right]_n$ is the first-order partial influence coefficient matrix influenced by platform rotation, and $\left[G_\phi^P\right]_n$ is the first-order partial influence coefficient matrix influenced by platform movement.

For each branch, the following equation can be attained:

$$\boldsymbol{V}_H = \left[G_\phi^H\right]\dot{\boldsymbol{\phi}} \tag{10}$$

where, $V_H = \begin{Bmatrix} \omega_h \\ V_P \end{Bmatrix} = \left\{\omega_{hx} \quad \omega_{hy} \quad \omega_{hz} \quad V_{Px} \quad V_{Py} \quad V_{Pz}\right\}^{\mathrm{T}}$,

$\dot{\boldsymbol{\phi}} = \{\dot{\phi}_1 \quad \dot{\phi}_2 \quad \ldots \quad \dot{\phi}_6\}$ is the generalized input velocity vector.

The above equation is the velocity of central point P in the moving platform coordinate system. Because there are 3 translational DOF in the mechanism, there is not rotation velocity in the moving platform. Consequently, the following equation can be obtained:

$$\boldsymbol{V}_H = \{0 \quad 0 \quad 0 \quad V_{Px} \quad V_{Py} \quad V_{Pz}\}^{\mathrm{T}} \tag{11}$$

When $\left[G_\phi^H\right]$ is not singularity，we have:

$$\dot{\boldsymbol{\phi}} = \left[G_\phi^H\right]^{-1} \boldsymbol{V}_H \tag{12}$$

The 3 translational DOF in the mechanism are active input, the last revolute pair in each branch is nominal pair (regarded as active input here). The 6 active input equations of the 3 branches taken from the upper equation are combined into a single expression, and then the following equation can be obtained:

$$\dot{\boldsymbol{q}} = \left[G_H^Q\right]\boldsymbol{V}_H = \left[\left[G_\phi^H\right]_{3:}^{-1(1)} \quad \left[G_\phi^H\right]_{3:}^{-1(2)} \quad \cdots \quad \left[G_\phi^H\right]_{6:}^{-1(3)}\right]^T \boldsymbol{V}_H \tag{13}$$

where, $\dot{q} = \{\dot{q}_1 \quad \dot{q}_2 \quad \dot{q}_3 \quad 0 \quad 0 \quad 0\}^T$ is the generalized input velocity, and $\left[G_\phi^H\right]_{\alpha:}^{-1(a)}$ is the α-th row of inverse matrix $\left[G_\phi^H\right]^{-1(a)}$ in the a-th branch.If $\left[G_H^Q\right]$ is not singularity and $\left[G_Q^H\right] = \left[G_H^Q\right]^{-1}$, we can get:

$$\boldsymbol{V}_H = \left[G_Q^H\right]\dot{\boldsymbol{q}} \tag{14}$$

Since there are 3 DOF in the mechanism and only three active inputs, we can obtain the matrix $\left[G_q^H\right]$, which consists of the first three columns form matrix $\left[G_Q^H\right]$. The matrix $\left[G_q^H\right]$ is the first order influenced coefficient matrix of parallel mechanism, which is the Jacobian matrix of the parallel mechanism. Correspondingly, we can obtain the velocity equation of input and output as follows:

$$\boldsymbol{V}_H = \left[G_q^H\right]\dot{\boldsymbol{q}} \tag{15}$$

In the above equation, $\boldsymbol{V}_H$ includes the angular velocity of moving platform and linear velocity $\boldsymbol{V}_P$ of a certain reference point P in the moving platform. We can get the velocity positive solution of the mechanism as follows:

$$\begin{aligned} \boldsymbol{\omega} &= \left[G_q^H\right]_{1\sim3:}\dot{\boldsymbol{q}} = [G']\dot{\boldsymbol{q}} \\ \boldsymbol{V}_P &= \left[G_q^H\right]_{4\sim6:}\dot{\boldsymbol{q}} = [G]_{3\times3}\dot{\boldsymbol{q}} \end{aligned} \tag{16}$$

where $[G]_{3\times3}$ is first order influenced coefficient matrix of the 3-UPU parallel mechanism. The inverse velocity of the mechanism can be expressed in the following equation:

$$\dot{\boldsymbol{q}} = [G]_{3\times3}^{-1}\boldsymbol{V}_P \tag{17}$$

5. THE PERFORMANCE ANALYSIS

A Performance analysis method

Based on kinematics influence coefficient method, the velocity of moving platform for parallel leg mechanism can be expressed as:

$$\boldsymbol{V}_P = [G]_{3\times3}\dot{\boldsymbol{q}} \tag{18}$$

In the above equation, $\boldsymbol{V}_P$ is the linear velocity of reference point P in the moving platform, $\dot{\boldsymbol{q}}$ is the input velocity, and $[G]_{3\times3}$ is a 3×3 first order influence coefficient matrix.

Take the derivative on both sides of the upper equation, equation (18) becomes:

$$\delta\boldsymbol{V}_P = [G]_{3\times3}\,\delta\dot{\boldsymbol{q}} \tag{19}$$

From equations (18) and (19), the following equation can be obtained：

$$\frac{\|\delta\boldsymbol{V}_P\|}{\|\boldsymbol{V}_P\|} \le \|\boldsymbol{G}_{3\times3}\|\|\boldsymbol{G}_{3\times3}^{-1}\| \le \frac{\|\delta\dot{\boldsymbol{q}}\|}{\|\dot{\boldsymbol{q}}\|} \tag{20}$$

The velocity performance index of the parallel mechanism can be defined as：

$$k_G = \|\boldsymbol{G}_{3\times3}\|\|\boldsymbol{G}_{3\times3}^{-1}\| \tag{21}$$

where, $\|\boldsymbol{X}\|$ is the Frobenius norm of the matrix $\boldsymbol{X}$.

Because $G_{3\times3}$ is not a constant matrix, its condition number k_G will change with the differences of positions and configurations of the parallel mechanism. Namely, different points in the workspace of parallel mechanism have different condition numbers. So we can not judge the velocity dexterity performance of the parallel mechanism by a variable. Consequently, we use the global conditional performance index to evaluate the velocity performance of parallel mechanism as follows:

$$\eta_G = \frac{\int_W \frac{1}{k_G} dW}{\int_W dW} \tag{22}$$

where, η_G is the velocity global performance index of parallel mechanism; W is the reachable workspace of parallel mechanism.

The mathematical meaning of the equation (22) is the average of the reciprocal of k_G in workspace, which is the average value of the performance index of all the points in reachable workspace. When $1 \le k_G < \infty$, we have $1 \ge \eta_G > 0$. Therefore, the larger the value of η_G is the higher mechanical dexterity and control accuracy are. Namely, mechanical velocity performance will be better.In addition, the kinematic equation of the parallel mechanism is $v = [J]\dot{\theta}$. The velocity performance index of parallel mechanism is defined as when the modulus of driven velocity vector $\dot{\theta}$ is unit 1, output is the extremum of the modulus of velocity vector v. Using this feature, not only the velocity and acceleration extremums, but also other extremums of various parallel mechanism end effecters can be obtained. The Lagrange equation is constructed as [12]:

$$L_v = \dot{\theta}^{\mathrm{T}} [J]^{\mathrm{T}} \dot{\theta} - \lambda_v \left(\dot{\theta}^{\mathrm{T}} \dot{\theta} - 1 \right) \tag{23}$$

The velocity extremums is the square root of the maximum and minimum eigenvalues of the matrix $[J]^{\mathrm{T}} [J]$, that is:

$$\begin{aligned} \left\| v_{\max} \right\| &= \sqrt{\lambda_{v\max}} = \sqrt{\max \lambda_{vi}} \\ \left\| v_{\min} \right\| &= \sqrt{\lambda_{v\min}} = \sqrt{\min \lambda_{vi}} \end{aligned} \tag{24}$$

B kinematic performance index map

Fig. 4 shows each kinematic performance index map of the 3-UPU mechanism when the circumcircle radiuses of upper and lower platforms are assigned different values whose range from 1 mm to 190 mm. It contains: the global performance index of the 3-UPU mechanism velocity, the global performance index maps of maximum and minimum values of the velocity. These figures show that the graphics are diagonal symmetry on the direction of 45°. Accordingly, the upper and lower platform of the 3-UPU parallel mechanism can be exchanged each other. This feature is consistent with the situation that leg mechanism acts as standing leg and swing leg in turn when walking robot is moving. It can be seen from Fig.4 (a) that when the circumcircle radius of moving platform ranges from 180 mm to190 mm, and the circumcircle radius of fixed platform ranges from 1 mm to 20 mm, (or reverse), the movement of mechanism in its workspace will have better isotropy. If the difference between the radius sizes of two platforms becomes small, the moving isotropy of the mechanism will become worse. Reference to Fig.4 (b) and Fig.4 (c): when the circumcircle radius of fixed platform changes from 20 mm to 110 mm, and the circumcircle radius of moving platform changes from 100 mm to 190 mm, the better performance indexes are showed. Therefore, we can draw a conclusion: when the difference between the radius sizes of two platforms is large, every performance index will be better.After modifying the structural parameters of the 3-UPU parallel mechanism and analyzing each kinematic performance index, we can obtain each performance map as shown in Fig. 5.

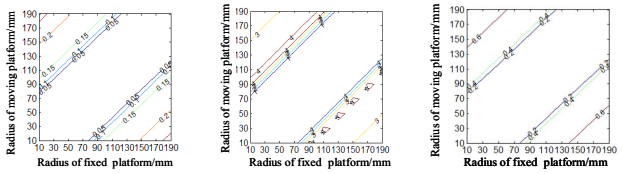


Fig. 4.The performance map of 3-UPU parallel mechanism (1)

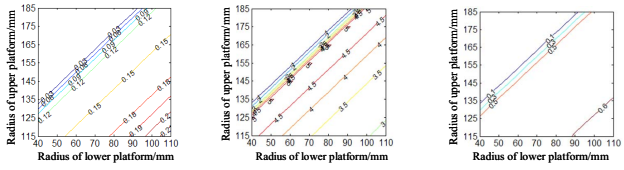


Fig. 5.The performance map of 3-UPU parallel mechanism (2)

The upper platform of the 3-UPU parallel mechanism acts as the hip connected with body, and the lower platform acts as the foot to bear the weight of the whole appliance. For the walking robot, the larger size of its hip will increase stiffness. Consequently, when the sizes of the upper platform meet the relevant sizes about the body linking, we should select larger sizes as much as possible. And the sizes of lower platform should be chosen by considering the load-bearing, walking stability and other factors during the robot walking.

6. CONCLUSION

1) According to the application of a typical 3-UPU parallel mechanism in a parallel walking robot, the inverse and forward kinematics of the 3-UPU parallel mechanism as standing leg is analyzed.2) Based on the first order kinematics influence coefficient matrix, the corresponding global velocity performance index of the 3-UPU parallel mechanism is analyzed.3) Present the analysis method of the kinematic performance index, and give the kinematic performance index map.

REFERENCE

[1] Z. Huang, L. F. Kong, and Y. F. Fang, Mechanism theory and control of parallel manipulators. Beijing, China:China Machine Press, 1997.

[2] R. Clavel, “Delta, A Fast Robot with Parallel Geometry”, Proc. Of the Int. Symp. On Industrial Rob., Switzerland, 1988, pp. 91-100.

[3] S. H. Li, Z. Huang, Y. Zhang, C. C. Yu, and W. H. Ding, “Design and analysis of 3-DOF micromanipulator driven by piezoelectric actuators”, Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, DETC 2008, pp. 871-878.

[4] R. Di Gregorio, V. Parenti-Castelli, “Mobility analysis of the 3-UPU parallel mechanism assembled for a pure translational motion”, IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM 1999, pp. 520-525.

[5] S. H. Li, Z. Huang, “Instantaneous kinematic characteristics of a special 3-UPU parallel manipulator”, Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference - DETC2005, 2005, pp. 691-697.

[6] T.S. Zhao, “Some Theoretical Issues on Analysis and Synthesis for Spatial Imperfect-DOF Parallel Robot”, Ph.D. dissertation, Yanshan University, Qinhuangdao, 2000, pp. 31-56.

[7] X. W. Kong, C. M. Gosselin, “Kinematics and singularity analysis of a novel type of 3-CRR 3-DOF translational parallel manipulator”, International Journal of Robotics Research, 2002, vol.21, no.9, pp. 791-798.

[8] Y. Yun, Y. M. Li, “Dynamics modeling for a novel 3-DOF dual parallel manipulator considering the flexibility of compliant components”, 2009 IEEE International Conference on Information and Automation, ICIA 2009, pp. 1410-1415.

[9] M. Ruggiu, “Kinematics analysis of the CUR translational manipulator”, Mechanism and Machine Theory, 2008, vol.43, no.9, pp. 1087-1098.

[10] L.W.Tsai, “Multi-degree-of-freedom mechanisms for machine tools and the like”, U.S. Patent Pending, No.08/415, 851,1995.

[11] H. B. Wang, Z. Y. Qi, Z. W. Hu, and Z. Huang, “Application of parallel leg mechanisms in quadruped/biped reconfigurable walking robot”, Journal of Mechanical Engineering, vol.45, no.8, 2009, pp. 24-30.

[12] X. J. Liu, “The relationships between the performance criteria and link lengths of the parallel manipulators and their design theory”, Ph.D. dissertation, Yanshan University, Qinhuangdao, 1999, pp. 11-25.

[13] Wu Z Q, Masaharu M. PID Type Fuzzy Controller and Parameters Adaptive Method[J]. Fuzzy Sets and Systems, 1996, 78(1): 23-35.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Research and Design of Campus Location Based Service System

Yugang Hu
Department Of Electromechanical Engineering
Changzhou Textile Garment Institute
Changzhou China
Huyugang80@163.com

***Abstract*—In order to solve the current smart phone location-based services LBS (Location Based Service) class of mobile applications in a small area of the environment can not be accurately provide location services issues, Presented to the Campus of LBS system design based on indoor and outdoor location. First introduced system architecture design, and detailed analysis of the Mobile GIS, GPS, system design based on the location of the fingerprint localization algorithm WiFi signal strength values of key technologies, Finally, test each module authentication system. Test results show that the system can achieve terminal location, indoor location services, campus navigation, map services and other functions, and convenient operation, with the feasibility and practicality.**

Keywords- location based service; Android; indoor-outdoor positioning; mobile GIS

I. INTRODUCTION

In recent years, the rapid development of mobile Internet technology, the growing popularity of Android smartphones based on Android platform developed various mobile applications emerging. Which is seen as mobile Internet "killer application" of LBS is to get an unprecedented development. LBS development so far, more people need to get accurate LBS services such as schools, hospitals and in such a small area of the environment. LBS provide services must be determined on the basis of the user's location, and this small area of the environment often contain both indoor and outdoor different environment. Outdoors, GPS provides very accurate location information; however, susceptible to blocking the satellite signal building, in the indoor environment can not provide highly accurate GPS positioning And WiFi, ZigBee, Bluetooth, infrared, ultrasound, radio frequency identification, ultra-wideband wireless location technology and rapid development, has become a strong complement to the GPS.

Therefore, this article campus this small area of the environment for the study of special functions, Design and implementation of a campus-based LBS system Android platform: In ArcGIS family of software produced by the campus map for the background; Outdoor environment by receiving GPS signals to locate; indoor environment without requiring the use of additional hardware devices premise, considering the wireless coverage area, by the impact of indoor environment, positioning accuracy requirements, etc., the use of location-based WiFi signal strength fingerprint location algorithm for positioning. The system can provide location-related information service for teachers, students and visitors.

II. SYSTEM DESIGN

A. System Requirements Analysis

The desired result is pursued in an outdoor environment, map service feature allows a user on a terminal device to view the campus environment map by zooming, moving up and down and other operating fully familiar with the campus environment. RTLS user location, query the optimal path travel destination from the current position and mark on the campus map; in an indoor environment, the user's selection displays the corresponding map of the indoor environment. Positioning, relative to the position of the user mark the indoor environment map on the indoor environment, and can acquire the service information corresponding to the position, for example, the empty classrooms information, program information and the like.

B. system design

According to the requirement analysis, the entire system is logically divided into client, server and database tier architecture.

Client: installed on Android smart phone system, complete the core functions and data of the front display, and the user is an important layer of interaction. System core functions include map service, outdoor GPS positioning, campus route guidance, indoor WiFi positioning, course information query, the query an empty classroom.

server: runs on the PC side, divided into GIS server, Web server and location server. The campus maps using ArcGIS Desktop software to create ArcGIS Server Posted on Web servers, and manage and update. When the Web server receives a request from a client to map the operation request, notice GIS server requires a call to the database according to the map data and corresponding geoprocessing tools to provide services. The location server is mainly used to run the algorithm, when a client receives a radio signal sent to call indoor location algorithm to determine the location of the terminal device, and transmits the location data to the client. Using a wireless network between the client and server for data transmission, communicates via standard HTTP protocol.

GIS Server: ArcGIS Server 10. 2
Web server: IIS 6. 0 (Internet Information Server6. 0)
Location Server: Apache Tomcat 7. 0. 47

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.55

Database: responsible for providing support to the service layer data. Clients use SQLite and a small amount of local data files stored; location server uses Mysql database to store the indoor location, location fingerprint database offline training stage, free classroom information and course information. GIS servers using Geodatabase storage campus map spatial data and attribute data.

III. CAMPUS MAP SYSTEM DESIGN

For a map of the campus environment design system it is to Campus GIS vector into the map and complete the publishing and management process, from the ArcGIS Desktop software to complete the acquisition of spatial data, edit, analyze, update and other operations, ArcGIS Server map services to achieve and network analysis service release, combined with ArcGIS for Android plug-in access to self-publish a map on the Android platform, access to map services and network analysis service.

A. Spatial Data Collection

Collection of spatial data is to convert data from different sources paper maps, remote sensing images, field observation data, text data, etc. can be received and processed into computer digital form. This map, including the following data: (1) map data, using a plane provided by the school campus map. (2) image maps, use GEtScreen software on GoogleEarth satellite image data interception campus. (3) the measured data, due to the limited facilities, a direct comparison of the map to obtain the map coordinates, road route length data measured by the existing map software ArcGIS Online, Google Map and the like.

B. Georeferencing

Campus flat map and satellite image data acquisition is free of any geographic data, and to use it will be carried out to give it the proper registration and geographic data. Here we must introduce the concept of spatial reference. The spatial reference includes X, Y, Z coordinates values and tolerance values and resolution values X, Y, Z and M values, using these attributes, you can determine the location of a feature on the earth. Common coordinate system includes geographic coordinates and projected coordinate system. This map, select a geographic coordinate system GCS_WGS_1984, the coordinate system is moving coordinate system platform used by GPS latitude and longitude coordinates of the information obtained by GPS coordinates are by this offer. ArcGIS in georeferencing using Geo-referencing tool bar, generally after selecting the coordinate system, add control points, check a few steps residuals, correction and resampling.

To reduce the impact caused the error, the control point should try to select the satellite image and relatively easy to distinguish fine features point or point at the edge of the image. And as far as possible in the area were uniform, full scale of the choice of site. Imaging features should be more changes in the larger region choose a few control points.

C. Spatial Data Editor

After completion of georeferencing, we need to build the campus diagram elements, the elements required to describe the campus manifested in the form of a layer, which is a vector of the process. According to the design requirements of the system, the spatial data is divided into five campus map layers, involving point features, line and area features three kinds of layer types, the completion of the feature vectors of different attributes. As shown in Table 1 below.

TABLE1 Spatial Data Layers

Layer Name	Type	Content
Buildings	Polygon features	School buildings, dormitories, administrative buildings etc
Playgrounds	Polygon features	Playground (basketball, tennis, etc.)
Roads	Line features	School road
CrossRoad	Point features	Intersection
Address	Point features	Address

Edited using Editor toolbar in ArcGIS spatial data sequentially for each feature layer editing. Note topology rules when the individual data elements for data editing.

D. Spatial Data Analysis

ArcGIS Network Analysis is used to simulate a variety of network problems to solve real world. Depending on the type of network problems, you can take a different modeling methods. For orienting network, usually geometric network analysis modeling approach. For non-directional network, usually the way modeling network dataset. Traffic Network is a non-directional networks for network modeling data set by the edges, junctions and turn feature composition. Proceed as follows: First, create and edit network datasets; path then centralized use ArcToolbox Network Analyst extension point to point in the network data analysis, get the shortest path.

E. Map Services And Network Analysis Service Published

On the ArcGIS Server server publishing services need to be installed before IIS, which is allowed in the network (including the Internet and LAN) publishing information on the Web server. After the campus map service, and network analysis service publishing success, enabling online access to the ArcGIS Online, different mobile terminals by IIS for calls. Android platform using ArcGIS for Android plug-in calls self-publish a map, perform the shortest path analysis operations.

IV. BASED ON THE LOCATION OF WIFI SIGNAL STRENGTH FINGERPRINT LOCATION

Complex indoor environments, the wireless signal propagation attenuation model can not accurately describe the relationship between a strong time-varying characteristics of the WiFi signal strength and distance. Due to the location of the fingerprint localization algorithm based WiFi signal strength value has better positioning robustness, so the system use it for indoor positioning.

Location fingerprinting positioning is based fingerprint terminal position location information, the location query fingerprint database, according to the corresponding matching algorithm to estimate the location of the terminal.

Location fingerprinting can get a variety, because RSSI easy to measure, which received widespread attention.

Location fingerprinting positioning is usually divided into two phases: Phase building a database offline and online positioning stage. Firstly, building a database offline stage geospatial coordinate a geographic map of the target environment corresponding map, and then the target area is divided into a number of collection points. Intensity level and the number of collection points can be set according to the environment and, in general, more intensive collection points, the more precise localization results. Multiple scans of each AP signal strength value for each collection point, the scan result "smoothing" process to remove some of the larger transition value, averaging the remaining values, the formation position RSSI fingerprint database. In this paper, a Gaussian filter method the signal "smoothing" process, the reason for taking the Gaussian filter method is due to the distribution of the normal distribution curve similar RSSI value. Based on engineering experience, we select the range of probability is greater than 0.6. After RSSI value ranges after Gaussian filtering process is [$\mu + 0.15\sigma$, $\mu + 3.09\sigma$].

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(RSSI_i - \frac{1}{n}\sum_{i=1}^{n} RSSI_i)^2} \quad (1)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} RSSI_i \quad (2)$$

According to the literature shows, AP access point to four the number of indoor environment location fingerprinting positioning accuracy to meet the requirements, so the location of the fingerprint database RSSI data table design <ID, X, Y, MAC1, RSSI1, MAC2, RSSI2 , MAC3, RSSI3, MAC4, RSSI4> form, where ID for each collection point number, X, Y for the horizontal and vertical coordinates of the point of collection, MAC1, MAC2, MAC3, MAC4 were four AP access points physical address, RSSI1 , RSSI2, RS-SI3, RSSI4 respectively corresponding to the signal strength average value is used for fingerprint feature location. Taking into account the layout AP campus buildings, most buildings can meet one floor four access points or more requirements, where no additional arrangement AP, choosing instead to average signal strength value maximum four existing AP conduct building a database. Measuring device uses radio signals Asus A45V, test software is specifically written a small program to acquire a wireless signal.

Online positioning stage, through the client scanning real-time signal strength value of the current position. In order to improve the accuracy of the data, will be set at 5 times the number of scans, then averaged value as a real-time signal strength of the AP. Deal will send real-time RSSI value to the positioning server, called by the location server nearest neighbor algorithm to match the location of the fingerprint database to identify the current location of the nearest collection point to estimate the location of the terminal.

V. CAMPUS LBS SYSTEM FUNCTION MODULE

A. Mapping Service And Campus Locations Inquiry

Map service module uses ArcGIS for Android SDK provides the core control MapView map to achieve load on the map, browse, zoom, pan and other operations. ArcGIS for Android in the Map-View as a map container used to present data map service. To display a map layer must be added to the MapView map container. First define MapView object, and then load the dynamic map layers onto campus MapView object by publishing a map service URL address corresponding to the REST interface to map services, achieve operational map services. The following are the key code to get a map service.

```
public class CampusMapActivity extends Activity {
private MapView mMapView;
private ArcGISDynamicMapServiceLayer mapLayer;
protected void onCreate(Bundle savedInstanceState){
super.onCreate(savedInstanceState) ;
setContentView(R.layout.campusmap) ;
mMapView=(MapView)findViewById(R.id.map) ;
mapLayer= new ArcGISDynamicMapServiceLayer( “
http://192.168.58.103:6080/arcgis /rest /services/Cam-
pusNew/MapServer” );
//Add campus map layers to MapView
MMapView.addLayer(mapLayer);}}
```

In addition to achieve operational map services outside interface also defines an EditText and a Button. Enter a place in EditText, click Button to complete the data analysis, obtain the corresponding location ID, and find the location of points in School location query.

B. GPS Positioning

Android platform, GPS positioning function is encapsulated in a LocationManager object. To use the GPS location services, you should first define LocationManager objects loc, open location services. After the service registered location services turned on by listeners LocationListener requestLocationUpdates method notifies listeners when positioning the current status or location changes. Custom Functions implemented within the listener function. Below are the Android platformGPS positioning on the call critical code.LocationManager loc;

```
Loc=(LocationManager)GetPositionActivity.this.getSys-
temService (Context LOCATION_SERVICE);
Loc. requestLocationUpdates (LocationManager.
GPS_PROVIDER, 0,0, new LocationListener () {
public void onStatusChanged (String provider, int status,
Bundle extras) {
..... / / GPS status change is triggered
}
public void onProviderEnabled (String provider) {
..... / / When the GPS is on the trigger
}
public void onProviderDisabled (String provider) {
..... / / GPS trigger disabled
}
public void onLocationChanged (Location location) {
Triggered ..... / / location change}});
```

C. Shortest Path Query

ArcGIS for Android execution path analysis can solve RoutingTask class method to carry out, will correspond to the first URL address network analysis on REST service interface is transmitted to RoutingTask object, and then call the method solve RoutingTask class and pass it RoutingParameters type parameters, you can find the shortest path. The following is to get the network analysis service, execute critical code shortest path query.

```
Try{
RoutingParameters rp = new RoutingParameters() ;
NAFeaturesAsFeature rfaf=newNAFeaturesAsFeature() ;
StopGraphic point1 = new StopGraphic(startPoint);
StopGraphic point2 = new StopGraphic( stopPoint) ;
Rfaf.setFeatures( new Graphic[]{point1,point2} ) ;
Rfaf.setCompressedRequest( true) ;
Rp.setStops( rfaf) ; rp.setOutSpatialReference(
MMapView.getSpatialReference( ) );RoutingTask rt = newRoutingTask(
" http://192.168.58.103:6080/arcgis/rest/services /
NewSchoolND/ NAServer /Route",null) ;
RoutingResult mResults = rt.solve( rp) ;
} catch( Exceptione) {
E.printStackTrace( ) ;
Looper. prepare( ) ;
Looper.loop( ) ;}
```

D. Indoor Location And Service

Online positioning stage, the client needs to scan the current signal strength value for the position. Android platform, this functionality is encapsulated in a WifiManager object. First, the object should be defined WifiManager wifimanager, permission to open the operation by WiFi getSystemService method; open after permission to establish the receiver broadcas-tReceiver, and override callbacks onReceive WiFi signal (); then () method to register the receiver broadcastReceiver by registerReceiver; Finally startScan () method to start scanning. Update function display interface after scanning results obtained when the trigger callback onReceive (), the callback function to send the scan to complete the location server, location server to obtain results that are returned. The following are key code gets neighboring AP signal strength on the Android platform.

```
Private WifiManager wifimanager;
private BroadcastReceiver broadcastReceiver;
broadcastReceiver = new BroadcastReceiver () {
public void onReceive (Context arg0, Intent arg1) {
List <ScanResult> wifilist=wifimanager. getScanResults
();
.....
}
}
Wifimanager=(WifiManager) getSystemService (Context.
WIFI_SERVICE);
/ / Register callback functions
registerReceiver (broadcastReceiver, new (IntentFilter
(Wifi-
Manager.SCAN_RESULTS_AVAILABLE_ACTION));
/ / Start scanning
wifimanager. startScan ();
```

for location and building select a user enters a display based on the corresponding indoor map, click on the "Get Current Position" button to start the scan, the scan results will be submitted to the location server, obtain a position location server returns information to inform the user fix red icon.

Location server estimated terminal position, according to new Date () method to get the current time, the query information table idle classrooms and curriculum information table, the query results are sent to the client.

VI. CONCLUSION

Through GPS positioning on the Android platform, based on research WiFi indoor positioning and Mobile GIS and other technology, design and implementation of a campus LBS system. Test results show that the various modules of the system can basically meet the basic requirements of teachers and students on the campus LBS services. But there are many places to be improved:

From the entire system, the data transmission client and server, rely on wireless communications network, a large amount of data, there will be the phenomenon can not interact with the network is not smooth.

Spatial data collection, lack of access to accurate data, professional equipment, data collection will be small errors.

Expansion modules, adding the surrounding recreational facilities inquiry, friends inquiries, etc., to make the system function more perfect.

REFERENCES

[1] SHEN Hongzhou,ZONG Qianjin,YUAN Qinjian.Implementation of Commerce Information Push Service Using Google C2DM[J], NewTechnology of Library and Information Service ,2012,28(6): 78-83.

[2] Google.AndroidCloudtoDeviceMessagingFramewo[EM/OL].[2012-04-09]. https://developers.Google.com/ android /C2DM. .

[3] ZHANG Jing,LIU Fu-ying. Study and Implementation of FoxNews-MID Based on Android C2DMK[J]. Computer Science,2011,38(10A):461-463.

[4] Aida Niknejad.A Quality Evaluation of an Android Smartphone Application[D]. Michigan:Eastern Michigan University, 2011,12-14

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Study on the IOT Architecture and Gateway Technology

Chang-le Zhong, Zhen Zhu, Ren-gen Huang
Foshan University
Foshan, China
E-mail: clzhong@fosu.edu.cn

***Abstract*—There are disadvantages in the practical application with Three-layer architecture of IOT (Internet of Things). In order to emphasize the level of IOT intelligent application, this paper introduces the five layers system architecture which can better interpret the meaning and features of the IOT, and discusses the gateway technology which connecting the sensing network and traditional communication network. According to the actual demand of hotel chain industry for improving guests' living environment, the paper also discusses the design of IOT application scheme with using the IOT gateway as a bridge. The scheme may effectively meet the service requirements of hotel chain industry.**

Keywords-Internet of things;system architecture;hotel chain;gateway technology;IOT application scheme

I. INTRODUCTION

Since the Internet of things (IOT) was proposed in 1999, its connotation has been in continuous development and expansion, but there are no uniform definition standards. The IOT concept broadly refers to RFID, infrared sensors, GPS, laser scanners and other information sensing devices, according to the agreed protocol, to achieve any time, any place, any object information exchange and communication in order to achieve intelligent identification, locate, track, monitor and manage a network[1]. The IOT has full perception, reliable transmission, intelligent processing and other features, and the IOT was making extensive use of, and made throughout the wisdom industry, wisdom agriculture, intelligent transportation, smart security, environmental protection, wisdom health care, government livelihood management, intelligent home, food safety and so on.

Although the IOT industry has been in rapid development in recent years, there is still no large-scale applications in reality. there is no uniform construction standards, norms things access and integration management platform[2]. The three-layer framework of IOT is widely regarded, namely, IOT is consisting of perception layer, network layer and application layer. Although the three-layer framework describes the architecture of the IOT from the technical level, but not fully shows the characteristics and connotation of the IOT. Now some applications require the closed-loop system, and the IOT is an open-loop ubiquitous network system, so its application and promotion still faces many difficulties and challenges.

In this paper, we put forward the five-layer framework of IOT that can better explain the characteristics and connotation of the Internet of things, based on the theoretical research of IOT and application in cross-regional chain hotel. In practical application, we realized the cross-regional hotel chain connectivity by using network gateway as a bridge in IOT.

II. THE FIVE-LAYER ARCHITECTURE OF IOT

Through analyzing the characteristics of the IOT[3], we put forward the five-layer framework of IOT. Namely, IOT is consisting of perception layer, network access layer, network layer, application support layer and presentation layer, as shown in figure 1.

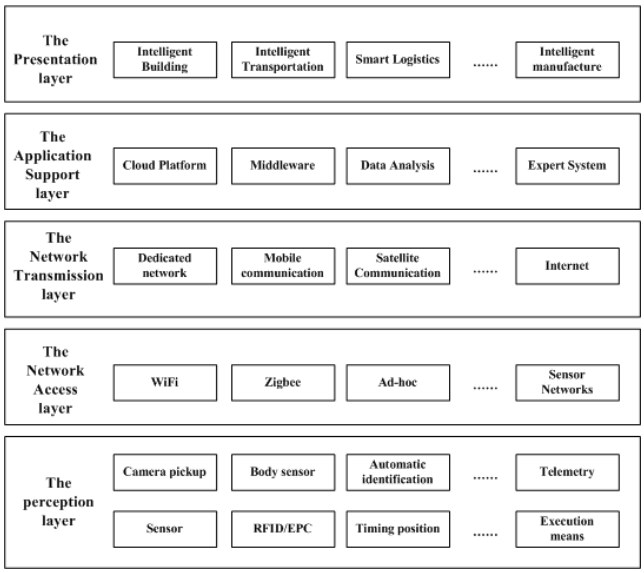


Figure 1. The five-layer architecture of IOT

(1) Perception layer of five-layer architecture: The perceptual layer is the foundation of IOT, is the interface between the layer of physical world and information world. It uses radio frequency identification technology, bar code technology, sensor technology, positioning technology, or other information sampling technology to complete the information collection, and with the help of controlling the objects of perception by the actuator, implement the infection control between the physical space and information space. Its main components include two-dimensional code label, code reader-writer, RFID tags and RFID reader-writer, cameras, and all kinds of sensors. So, The IOT perception layer has the main functions of information perception and original data collection, necessary auxiliary complete downward at the end of the control object. Therefore, the main function of perception layer of IOT is information and

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.56

data collection, when necessary, assist to complete the control objects of perception.

(2) Network access layer of five-layer architecture: The network access layer is mainly composed of the base station node and the network access gateway, complete the network control and the data fusion of each node in the perception layer, or complete to forward the information from the above layers (The network transmission layer or the application layer). When the perception layer's nodes complete networking, the perception layer's nodes need to upload data, and send the data to the base station node. The base station node will receive the data, and complete the connection with the network transmission layer by the access gateway. When the application layer and the network layer needs to downlink data, the base station node sends data to each node in the perception layer after the network access gateway receiving the data from the network transmission layer, then complete the forwarding information and interaction between the perception layer and the network transmission layer. The current access methods in the network access layer mainly include WIFI, Ad hoc, Mesh, ZIGBEE, industrial bus, realize to collect the information by various cognitive tools, or to preliminary process and network access.

(3) Network transmission layer of five-layer architecture: The network transmission layer is mainly used to realize the transmission and exchange of information, provide the basis transmission network for the necessary of applications and services within a wide range, including the satellite communication network, the mobile communication network, the optical fiber communication network and the local independent private network and so on. It is a problem in the network layer that the neutral access and seamless integration between different network and means of communication, and how to form the transmission and exchange capacity with end-to-end .

(4) Application support layer of five-layer architecture: With the support of the information technology with the cloud computing technology, middleware technology, database technology, expert system and so on, the application support layer complete public intelligent analysis and storage of data information, realize information processing, and all kinds of intelligent application sharing and exchanging.

(5) Application presentation layer of five-layer architecture: The application presentation layer's task is the development of a variety of applications of IOT base on the data processing of the application support layer, and uses the technology with multimedia, virtual reality, human-computer interface to build the interface of intelligent application between the IOT and the user, implement present and application of all kinds of intelligent information.

III. THE GATEWAY TECHNOLOGY OF IOT

The gateway is a network to another network "mark", and the IOT gateway is a connecting link between the sensor network and the traditional communication network, and it can store and convert the interaction data between the network and the traditional communication network. The gateway is between the perception layer, network layer and the network access layer in the five layers of IOT architecture, and the IOT will be able to integrate a variety of access methods, to meet the convergence and access requirements for the local short distance communication, to link to the public transport network and complete the forwarding, controlling, signaling and encoding and decoding functions. At the same time, the gateway is protocol converter, to achieve protocol conversion between two different networks, and then a packet format can be converted to another packet format, and has the function with safety protection and prevent outside intrusion. So, the IOT gateway will have access ability between the WAN, management ability, protocol translation ability, and other major functions[4].

A. The IOT gateway hierarchy

IOT gateway supports a variety of communication protocols and data types between the various sensors, which can realize the conversion of data format which communicated between a variety of sensors, to unified the uploaded data formats. At the same time, the acquisition or control command which reach the perception network are mapped to produce messages that meet specific device communication protocol[5]. The basic structure of the IOT gateway as show in Figure 2, including application layer, network layer, analysis layer (protocol conversion and protocol adaptation layer), perception layer.

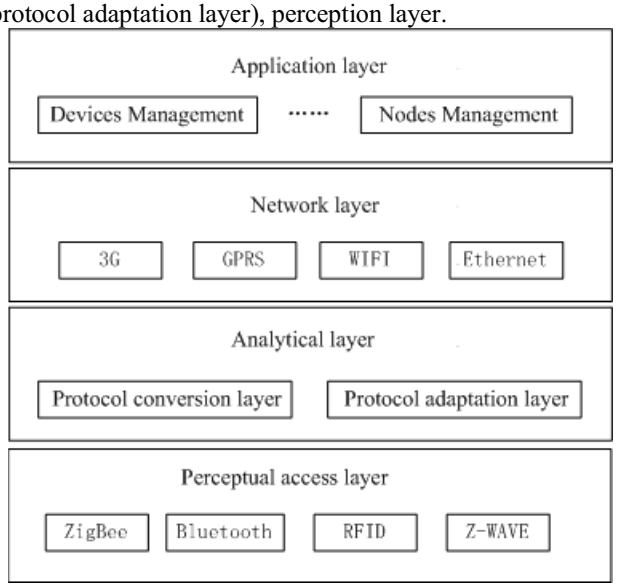


Figure 2. The IOT Gateway Hierarchy

(1) The application layer: The application layer will realize automatic management of sensing device, and the management of each sub network.

(2) The network layer: The network layer provides a variety of channel access communication network interface. For the mobile environment or non fixed environment of the network, you can use a variety of access methods. For the specific network environment of the network, can be used in single access mode. The network layer includes a variety of

communication network and Internet to form network, which is the current mainstream of communication network, such as 2g, 3g, 4g network or a computer Internet, etc.

(3) The analytical layer: The analytical layer will implement the standardization of protocol conversion and data format analysis, including the protocol adapter and protocol conversion module.

The protocol adapter module defines a interface to access standard, ensuring different access layer protocols can become a unified data format and signaling.

The protocol conversion module will be unified packaging the uploaded standard data from the protocol adapter, unpack the data from the network layer into the standard format. And provide the protocol conversion from perception to communication network , namely to implement ZigBee protocol to TCP/IP protocol conversion.

(4) The perceptual access layer: The perceptual access layer will complete the network control and the physical access for nodes, and match a variety of sensor network technology to realize different perception network protocol access.

B. The gateway hardware structure

The IOT gateway is a bridge connecting perception network and access network, it can support different types of sensor nodes (such as ZigBee, 6LoWPAN, RS485, CAN) and the way of access (such as cable, WLAN, GPRS, 3G), and provide a unified data format for middleware or application, in order to shield the different sensor network and the access network, make applications only need to pay attention to in the application environment of data processing.

This paper adopted the modular design concept and embedded system technology to design the IOT gateway, the structure of the IOT gateway is shown in figure 3. The processor module is the core module of the gateway, which implements the protocol conversion, management, security and other aspects of data processing and storage. The zigBee module realize the collection of physical world data or together, can be the convergence of sensor network nodes, the RFID reader, video collection equipment, GPS, etc. Through the network access module, the gateway will access WAN by the way including cable (Ethernet, ADSL, FTT), wireless (WLAN, GPRS, 3G, satellite).

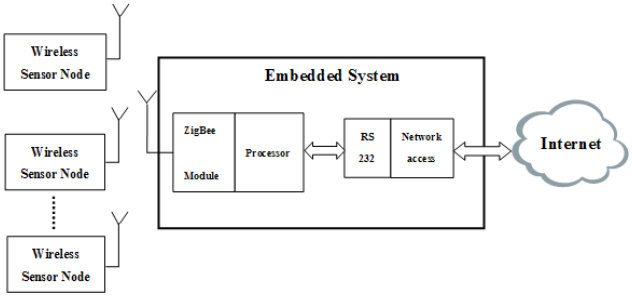


Figure 3. The gateway hardware structure

IV. THE DESIGN OF THE IOT GATEWAY OF CROSS-REGIONAL CHAIN HOTELS

For chain hotels, to build an efficient and stable cross-regional network management architecture is very important. And how to make the devices (such as air conditioning, television, access control, lighting, etc.) to interconnect, effective sharing of resources and information to facilitate the stay guests, is an important indicator of the quality of service of a hotel. With the development of IOT technology, the IOT gateway technology become more mature, and the equipment in the hotel room connectivity problem will be solved.

The IOT gateway, which is the core equipment of the construction and intelligent of hotel rooms, allows multiple intelligent devices interoperability in the room, forms a local area network, shares resources and information between devices. In addition, the gateway also plays the internal network and the external network interface communication role, providing access and control functions for all kinds of value-added internal network service, which makes the internal rooms network become an extension of the communication network, communication network the internal network to the wide world.

A. application architecture

Based on the study of design for the smart home system [6][7], we put forward a typical cross-regional chain hotels IOT application architecture, as shown in figure 4.

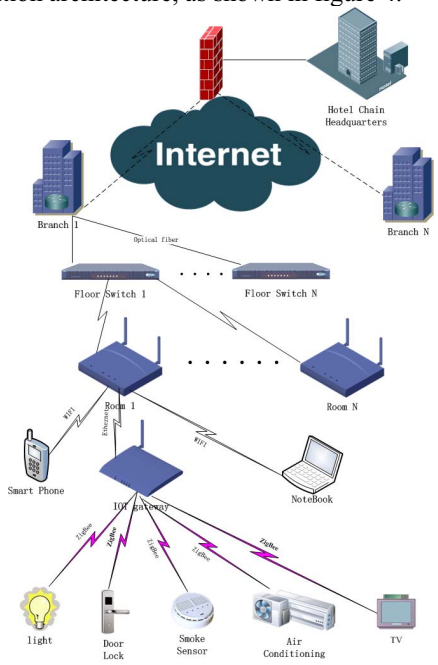


Figure 4. The IOT structure of Cross-regional hotel chain

In the hotel guest room, for TV, air conditioning, access control, smoke detectors, lighting and other equipment, the system will be build their own subnet system respectively, and make the equipment of different protocols or subsystem to communicate with each other through the gateway, and

users only need to operate the gateway to control all intelligent devices connected to the gateway. In addition, the gateway is also integrated with internet access, can carry on the wide-area interconnection with the outside world, the device within the hotel room can be operated and control in any place of the world, then greatly enhance the convenience and applicability.

B. the design of the gateway hardware[8]

In this paper, we use STR912 ARM9 chip which integrated Ethemet (MAC) interface to design the IOT gateway, it mainly includes the embedded ARM processor module, Zigbee protocol module and network access module, the structure is shown in figure 5.

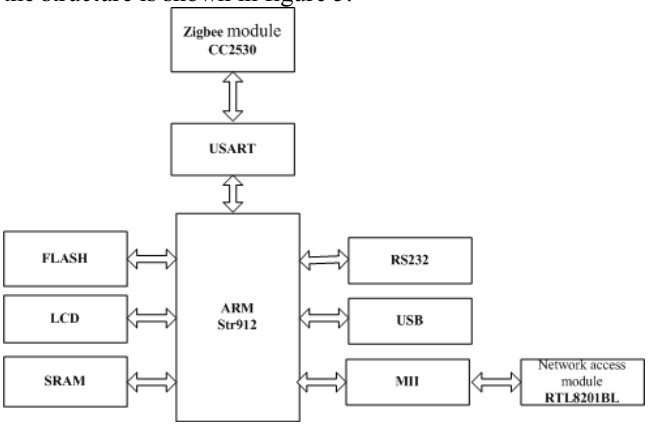


Figure 5. ARM-based processing module structure

The processor module is the core of the gateway, the gateway is responsible for the entire device control, task allocation and scheduling, data integration and so on transmission. Its core chip using ST's STR912FW44. STR912FW44 is one of the latest series of industrial-grade single-chip ARM9 microcontrollers, based on ARM966E-S core, on-chip comes with FLASH, USB, CAN, SPI, Ethernet and other peripherals, with tightly coupled 512KB FLASH memory and 96KB of SDRAM. STR912 works with stable performance, strong compatibility, strong scalability, maintenance reliability, save large data capacity, long time and so on.

The zigbee protocol module uses the CC2530 of the Texas Instruments (TI) , for data communication between the gateway and the sensor nodes, the node is responsible for receiving data. The CC2530 combines a high-performance 2.4GHzDSSS (direct sequence spread spectrum) RF transceiver and a high-performance low-power 8051 microcontroller, used to build the network node with low price, at the same time, it integrated IEEE802 .15.4 standard 2.4GHz band RF transceiver in a single chip, with excellent radio reception sensitivity and immunity.

The network access module is used to the gateway access wan, because STR912FW44 processor module is included within the Ethernet MAC and MII interface, when connected to the Ethernet network , the access module only need to add an Ethernet physical layer (PHY) chip RTL8201BL.

As can be seen from the above analysis, the processor module STR912FW44 mainly receive the data from the zigbee wireless sensor network and realize the data storage and process, and then send the processed data sent to the Internet until the headquarters center server through the network access module(RTL8201BL). At the same time, for some control commands sent by the server center headquarters to process, and transmitt to the nodes of the sensing network. In addition, the gateway can also include a LCD display, buttons and other functional components.

V. CONCLUSION

The concept of IOT is for a long time, but the specific implementation and composition framework of IOT have not formed a unified opinion. In this paper, on the basis of analysis of practical applications of IOT in life, we analysis lack of research on the three layer system structure, put forward five layers of IOT application system frame of reference, it can better explain the characteristics and meaning of IOT. At the same time, through the analysis of demand, we put forward the solving scheme of cross regional hotel chain of the IOT combined with the technical characteristics of the IOT application, and use the IOT gateway as a bridge to realize exchanges of the hotel rooms information and communication of different equipment, realize the connectivity of the IOT with the Internet, and to meet the needs of the cross-regional chain hotel services.

REFERENCES

[1] Zhang Mingjie, Han Jianting, Hu Bingsong, Liu Wenchao.Building Home Application System of Internet of Things with Home Gateway.TELECOMMUNICATIONS SCIENCE, 2010(4),P44-47

[2] ZHANG Wei,ZHANG Zhe. Access Technique of the Internet of Things Gateway. Journal of Nanyang Normal University,2012(12),P68-70

[3] Hai-tao Zhang, Yong-kui Zhang. Architecture and Core Technologies of Internet of Things. Journal of Changchun University of Technology (Natural Science Edition), 2012(2), P176-181

[4] Huang Haikun,Deng Jiajia. Discussion on the Technology and Application of IOT Gateway. TELECOMMUNICATIONS SCIENCE, 2010(4), P20-24

[5] Junhai luo,Yingbin Zhou, Xiaobo Deng. Design for Gateway in Internet of Things. TELECOMMUNICATIONS SCIENCE. 2011(2), P105-110

[6] Jun Hou, Cheng-dong Wu, Zhong-jia Yuan, Yun Zhou, Yun-zhou Zhang. Research of Intelligent Home Security Surveillance System Based on ZigBee. Mechanical & Electrical Engineering Magazine, 2009(1), P67-70

[7] Zhong-liang Nan, Guo-xin Sun. Design of Smart Home System Based on ZigBee. Electronic Design Engineering, 2010(7),P117-119

[8] GAN Yong,WANG Hua,CHANG Ya-jun,WANG Jun. Design of Zigbee Gateway System Based on ARM. Communications Technology, 2009(1),P199-201

Simulation Study on Multi-lane Traffic Flow under Right-most Overtaking Rule Based on Driving Security Determination and Assistance Overtaking System and Intelligent System

Hongxia Wang, Wenkai Guan, Yue Yu
School of Computer Science & Technology
Wuhan University of Technology
Wuhan, Hubei, China
e-mail: whx_green@163.com;
passionguan@whut.edu.cn;
wo4li2wang@gmail.com;

Dongfei Liu
School of Statistic
Wuhan University of Technology
Wuhan, Hubei, China
e-mail: 1054467130@qq.com

Qiyu Liang
School of Transportation
Wuhan University of Technology
Wuhan, Hubei, China
e-mail: 975085417@qq.com

Yongsheng Yu
State Key Laboratory of Silicate Materials for Architectures
Wuhan University of Technology
Wuhan, Hubei, China
e-mail: yongshengyu@whut.edu.cn

***Abstract*—This paper proposes two systems, Driving Security Determination and Assistance Overtaking System and Intelligent System, to analyze the performance of the right-most rule on multi-lane traffic flow. Driving Security Determination and Assistance Overtaking System(DSDAOS) is established to analyze the relationship between security, speed limit and traffic flow while Intelligent System(IS) is theoretically designed on the basis of DSDAOS and it has better impact on promoting traffic flow from aspects of speed limit and lane conditions when compared to DSDAOS. DSDAOS is a system that utilizes computer to produce vehicles in the simulation, give some properties to them which are made to run in lanes according to certain rules, and finally output security level and relational traffic flow. *IS* is a system in which computers use information-gathering device to automatically gather information, then deal with the information and get relating response which refers to the function of shifting speed limit of lanes and changing lanes to promote traffic flow. All in all, IS makes great difference to promote use ratio and carrying capacity of lanes. What is more important is that Intelligent System is a circulating system with high independence.**

Keywords-traffic flow; overtaking rule; driving security determination and assistance overtaking system; intelligent system

I. INTRODUCTION

Traffic jams and the trends of multi-lane traffic promoted the research and development of traffic flow theories[1]. In efforts to analyze the multi-lane traffic flow, mathematical models are currently used. A possible classification is a division into microscopic, macroscopic, and kinetic models[2]. In microscopic models, we consider the dynamics of each single car. But the macroscopic models describe quantities such as traffic flow velocity and traffic density. As to the kinetic models, they deal with probability distributions.

Intelligent Transportation System(ITS) are advanced technologies which aim to provide innovative services relating to different models of transport and traffic management[3]. Enhanced vehicle flow, reduction and eventually elimination of traffic congestion have been attractive issues for the past two decades in the Intelligent Transportation Systems(ITS) domain[4]. Representative solutions to this are Advanced Driver Assistance Systems(ADAS) and Adaptive Cruise Control(ACC) and etc.

Goals which are reached in this paper:

- The relationship between the security, speed limit and traffic flow was analyzed through the establishment of Driving Security Determination and Assistance Overtaking System(DSDAOS): we studied the relation between security and traffic flow under given speed limit; we also analyzed the relation between speed limit and traffic flow when a security level is given.
- Intelligent System (IS) was designed and built on the basis of DSDAOS, the change in mathematical model when compared with DSDAOS was discussed under the complete control of intelligent system, which focus on the Intelligent System's promotion to traffic flow under the rule of "driving on the right and overtaking from the left".

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.57

II. MODEL 1: DSDAOS

A. Assumptions

- Vehicles which are inputting to computers remain in the same lane at a uniform speed without overtaking other vehicles.
- The time when vehicles accelerate to shift lanes and decelerate to shift lanes can be neglected.
- When simulating the driving conditions of freeways, there is assumed to be no traffic accident.
- Both systems and models are built on the assumption that the weather is fine, the impact of the weather on systems and models is not taken into consideration.

B. The Setting of Security Level of DSDAOS

S : Distance from the tailstock of the vehicle ahead to the headstock of the vehicle behind before the overtake;

V : driving speed in the overtake lane while overtaking;

S' : Distance from the tailstock of the "vehicle behind to the headstock of the "vehicle ahead" after the overtake.

In order to classify security level, three critical values should be identified, as follows in table 1:

TABLE I. CRITICAL POINTS OF SAFETY DISTANCE S [5]

Critical point	response time of drives t_1	response time of vehicle braking t_2	error increment Δt	safety distance S
1	1s	0.9s	0.5s	S1
2	0.7s	0.6s	0.5s	S2
3	0.3	0.2	0.5s	S3

Each of the three parameters, named S , V , S' , is divided into four degree: Very dangerous, Dangerous, Safe, Absolutely safe[5]. The classification of the security level is as follows in table 2:

TABLE II. CLASSIFICATION ACCORDING TO THE RULE[5]

Indexes / Security Level	S	V	S'
1	$[S_0,S_1]$	$[V_3,V_4]$	$[S'_0, S'_1]$
2	$[S_1,S_2]$	$[V_2,V_3]$	$[S'_1, S'_2]$
3	$[S_2,S_3]$	$[V_1,V_2]$	$[S'_2, S'_3]$
4	$[S_3,S_4]$	$[V_0,V_1]$	$[S'_3, S'_4]$

Note:1 represents "Very dangerous",2 represents "Dangerous",3 represents "Safe",4 represents "Absolutely Safe".

C. Construction of DSDAOS

The construction of the whole system is as follows in the Figure1.

1) Input section of the system

According to existing knowledge, it can be noted that the number of vehicle through a certain cross section abides by Poisson distribution. Furthermore, max speed and minor speed varies from lane to lane in freeways. Finally, the type of vehicles as well as length, driving speed passenger car unit of them is various. For the sake of convenience, with the combination to the urban traffic regulations, the type of vehicles are simplified as follows in the table 3.

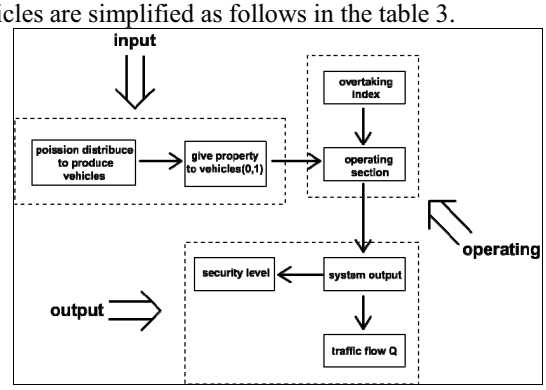


Figure1. The construction of the System

TABLE III. CONVERSION OF LENGTH AND PASSENGER CAR UNIT OF VARIOUS TYPES OF VEHICLES

Type of vehicle	Length of vehicle/m	passenger car unit of vehicle/pcu*h^{-1}
small-size	5	1
medium-size	12	2
large-size	18	3
extra-large-size	20-40	5

2) Vehicles Produced in the Simulation

A given strength λ combined with proposition to number of various vehicles and physical truth is defined as 10:5:2:1. Computers are utilized to perform Poisson simulation and produce vehicles $x_1, x_2, x_3 \cdots$

3) Giving Properties to Vehicles

Computers are utilized to give relating property $x(v,l)$ to vehicles produced with combination to design speed of vehicles and limit speed of lanes, for example, the minor speed is limited to 60 km/h in through lane ,the max speed 180 km/h , the length of small-size vehicles are automatically given to five meters, at the same time, a speed between 60 km/h and 180 km/h produced by computers by way of uniform distribution is given to the vehicle.

4) Operation Section of the System

According to the assumption[1,2,3].For the convenience of the simulation, some limits are added to vehicles which run in the lanes:

- Vehicles input from computers remain in the same lane at a uniform speed without overtaking other vehicles.
- Vehicles run at a uniform speed while overtaking other vehicles and the time to accelerate to shift lanes and decelerate to shift lanes can be neglected.

- The rule that overtake from the left must be abided by when overtaking is performed.

Given overtaking indexes (S, υ, S'), at given intervals, more vehicles are produced to simulate the reality with increase in input strength(λ).

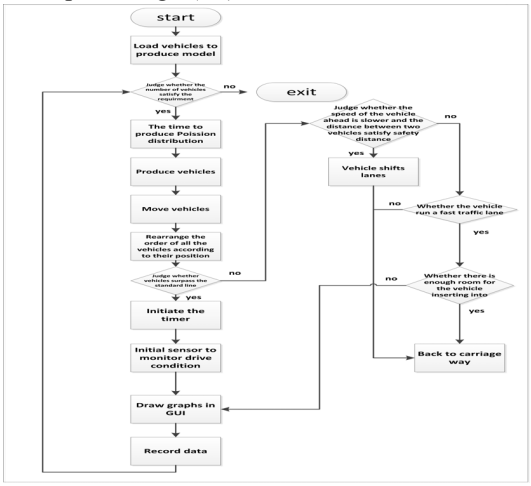


Figure2. Procedure of the System

5) Results and Analysis od the Simulation of Four-lane DSDAOS

The result is reached as follows:

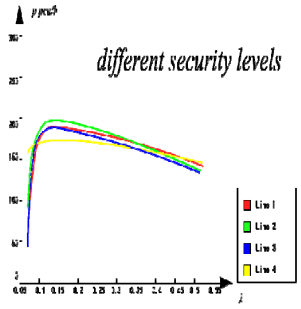


Figure3. Relations between P and λ under different security levels

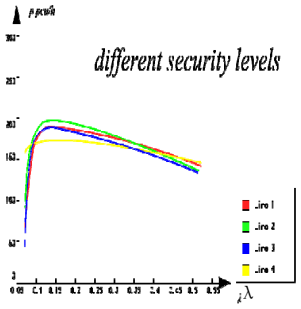


Figure 4. Relations between P and under different max speeds

In Figure 3, Line 2 stands for the set range when the rule is applied under current China's traffic regulations. According to the trend revealed by four lines, when traffic flow is low, the traffic flow output will increase in the first period; while the traffic flow input reaches its peak, the traffic flow output will decrease because lanes are crowded. From Line 1to Line 4, with security level decreasing, the traffic flow output will firstly increase and then decrease, and the traffic flow output reaches peak in Line 2.

In Figure 4, the trend revealed by Line 1can be applied to the rule. According to the trend revealed by four lines, the traffic flow output firstly increases but then decreases. From Line 1to Line four, with max speed decreasing, the max traffic flow output continues to decrease and reaches peak in Line 1.

III. MODEL 2: INTELLIGENT SYSTEM

A. Relationship between the Speed of Vehicles and the Traffic Flow

After looking for scientific literatures, the conclusion can be reached that traffic flow is related to the speed of vehicles in the street, ideal condition is as follows in the Figure 5.

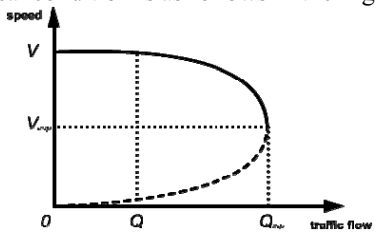


Figure 5. Relations between V and Q under ideal condition

B. The Structure of IS

IS is composed of three parts: information-gathering device, information-processing device and response-outputting device. The structure of *IS* is shown on Fig.6.

Before the system operates, a initial phase is given to it: optimal overtaking indexes (s, v, s'), intervals to update time T, road speed limit $v_{\max}$.

1) Information-gathering Device

Sensors records the number of passing vehicles $N(t)$, the time when vehicles pass the sensor t_i and the information about the lane where vehicles run.

2) Information-processing Device

First of all, Information-gathering Device passes the information gathered to information-processing device and then calculate the traffic flow in the duration by using the relation between the number of passing vehicles ($N(t)$) and time (t_i).

$$Q = \frac{N(t)}{T}, \Sigma t_i = T \qquad (1)$$

Then, make a comparison between Q_{cap} and Q: when $Q > Q_{cap}$, change the property of lanes, when $Q < Q_{cap}$, the optimal speed is reached with combination to the relation between speed v and traffic flow Q.

3) Response-outputting Device

According to the results the above mentions, the initial phase should be reset. Then, the system processes next procedure and will continue to work like this.

4) Analysis of the Results

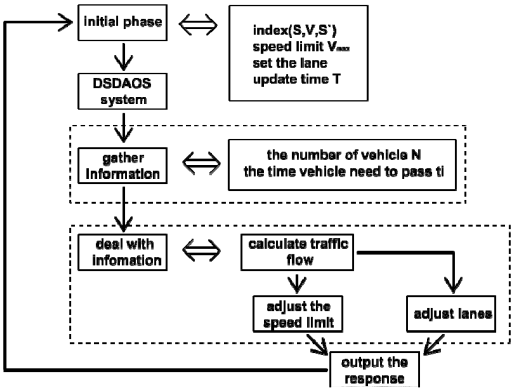


Figure 6. The Compose of IS

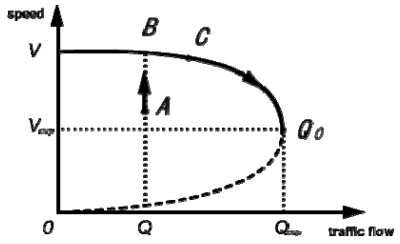


Figure 7. Traffic Conditions with *IS* and without *IS*

According to Figure 7, points on and inside the parabola represent all the possible conditions a part of a lane is possessed of under certain circumstances.

- Point A which firstly exits inside the parabola shifts to Point B after the adjustment of intelligent system, thus increasing the max speed of lanes with the same traffic flow and save time.
- Point C which firstly exits on the parabola, which means traffic flow has saturated with current speed, however, traffic flow approaches Q_0 along the parabola to get larger traffic flow after IS decreases the speed limit.
- At the same time, in terms of adjusting lanes:

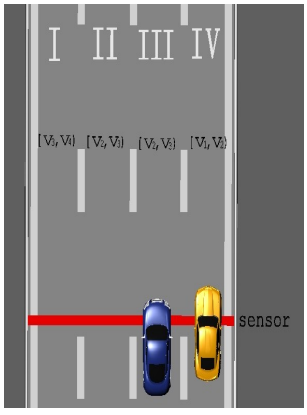


Figure 8. Before the adjustment of IS

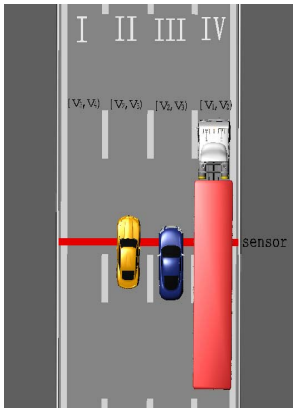


Figure 9. After the adjustment of IS

Figure 8 indicates that traffic flow in the third lane does not reach Q_{cap}, and the second lane is the overtaking lane, the third and fourth lanes are carriage ways;

Figure 9 indicates that traffic flow in the third lane reaches Q_{cap}, and the second, third and fourth lanes are carriage ways, thus making sure that use ratio of lanes is improved with the same traffic flow.

IV. CONCLUSION

According to the results of the simulation study, we can find out that with the analysis of DSDAOS, the traffic flow, which employ the Right-Most Overtaking Rule, will increase in the first period when traffic flow is low; While the traffic flow is heavy and lanes are crowded, this rule will decrease the traffic flow of the multiple lanes. We also notice that with security level decreasing, the traffic flow output by the DSDAOS will firstly increase and then decrease.

However, the rule as stated above relies upon human judgment for compliance. If vehicle transportation on the same roadway was fully under the control of an intelligent system, it can improve the traffic flow from aspects of speed limit and lane conditions: 1). IS can increase the max speed of lanes with the same traffic flow and save the travel time when compared to DSDAOS; 2). When traffic flow has saturated with current speed, it can decrease the speed limit thus to get larger traffic flow.

The IS just makes improvements from aspects of speed limit and lane conditions and has no chance to explore improvements from other aspects. In the future, there will be more ways to measure the degree of promotion of rules to traffic flow in order to change the rule to get the optimal traffic flow.

ACKNOWLEDGMENT

The paper is supported by the Natural Science Funds of Hubei Province (Grant No. 2013CFB351), and the Fundamental Research Funds for the Central University (Grant No. 2014-IV-105).

REFERENCES

[1] Zhou Fang, Ruolan Li, Fuzhou Li, Ziqin Zhou, "Modeling Multi-lane Traffic Flow under Different Overtaking Rules Based on Cellular Automaton[C]", Computer Science & Education, Vancouver, Canada, pp. 647–653, August 2014.

[2] Lubor Buric, Vladimir Janovsky, "A traffic flow model with overtaking as a Filippov systems[J]", Journal of Computational and Aplied Mathematics, vol. 254, 2013, pp. 55–64.

[3] Xinping Yan, Hui Zhang, Chazhong Wu, "Research and Development of Intelligent Transportation Systems[C]", Distributed Computing and Applications to Business, Engineering & Science, Guilinr, China, pp. 321–327, July 2012.

[4] Aleksandar Kostikj, Milan Kjosevski and Ljupcho Kocarev, "Harmonized Traffic Stream in Urban Environment Based on Adaptive Stop & Go Cruise Control and its Impact on Traffic Flow[C]", Vehicular Electronics and Safety, Istanbul, Turkey, pp. 140–145, July 2012.

[5] Leonard Evans, "Traffic Safety and the Driver[M]", Science Serving Society.1991.

Real-time Calculation of Road Traffic Saturation Based on Big Data Storage and Computing

Youwei Yuan, Linliang He, Wanqing Li, Lamei Yan
School of Computer Science and Technology
Hangzhou Dianzi University
Xiasha, Hangzhou, Zhejiang, P.R China
e-mail: yyw@hdu.edu.cn

M.Mat Deris
Faculty of Computer Science and Information Technology
University Tun Hussein
Onn Malaysia, Johor, Malaysia

***Abstract*—Road traffic saturation data is streaming at unprecedented speed and must be dealt with in a timely manner. In the paper, we put forward a real-time calculation method of road traffic saturation based on big data storage and computing. We will calculate the road traffic saturation data via adjacent traffic checkpoint while the computing and storage of road saturation based on Spark big data processing engine of big data and the final result data storage to the non-relational database HBase. The experiment results show that the performance of our method is superior to the original methods mainly embodied in the advantages of precision and real-time.**

Keywords-Big data; Storage; Traffic checkpoint; Road saturation

I. INTRODUCTION

With the dawn of big data, more and more data processing engine had emerged, and Spark is now very efficient big data processing engine. The definition of big data is increasingly being used to refer to the challenges and advantages derived from collecting and processing vast amounts of data[1], but it is defined as the processing capacity of the system exceeds a given number of data from the form[2]. In this era of big data, more and more research field are closely linked with big data. With more than 950 million users, Facebook is collecting500+terabytes of new data ingested into the databases every day [3]. The complexity of big data storage depends on the optimization of the storage of historical data and real-time data[4-6].

With the rapid development of vehicles, traffic data has greatly increased. Traditional road traffic saturation method is artificial calculation. The road saturation is the original data stored in the relational database, and then operate the data in the database, and finally the results of the data stored in the relational database while greatly reduce the efficiency of data storage. So the result is not real time and the process of computing is very slow .It is very necessary to develop a new method for real time computing and storage road traffic saturation efficiently.

This paper proposes a solution to the computing and storage of road saturation method in big data background. This approach comprises several key steps: the traffic data acquisition and storage to non-relational database HBase. The structure of the paper is as follows: The theoretical are described in Section 1. Introduces the specific steps of data processing engine Spark technology and algorithm of adjacent traffic checkpoint is presented in Section 2. The experiment results are discussed in Section 3. The conclusions are presented in Section 4.

II. METHODS AND TECHNIQUES

A. *Road saturation based on Spark big data processing engine*

Spark is an open source project from Berkeley University AMPLab development free, it is big data processing engine based on memory calculation[7-8]. Compared to other big data processing engine Hadoop, it is more suitable for the requirements of the scene of the time, because it overcomes the MapReduce model with high delay fatal weakness[9-10].

The flow chart shown in figure 1 to calculate the road saturation data processing engine based on Spark. We firstly get the vehicle information datas of non-relational database in HBase while the data stream is divided into a number of suitable size flexible distributed data set RDD. Then we will calculate the value of every road saturation at the set time range .Finally all road saturation value of the data will be stored in a HBase.

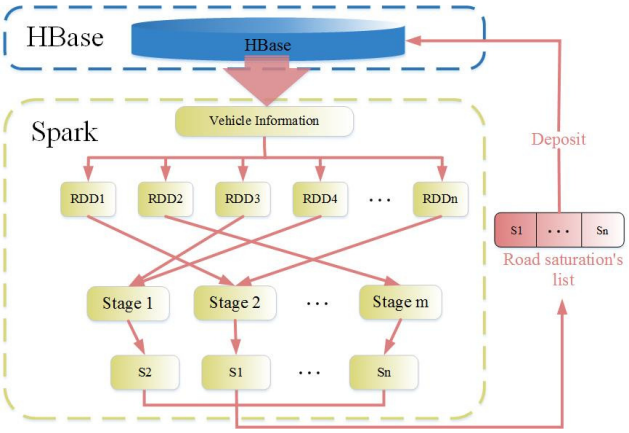


Figure 1. Flow chart of road saturation based on Spark big data processinengine

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.58

B. *Real time road saturation calculation based on Spark Streaming*

This method is a combination of a sub frame Spark Streaming data processing engine Spark to real-time calculation of road saturation, while can improve the performance of data analysis algorithm[11-12]. Our real time road saturation calculation framework is shown in Fig. 2.

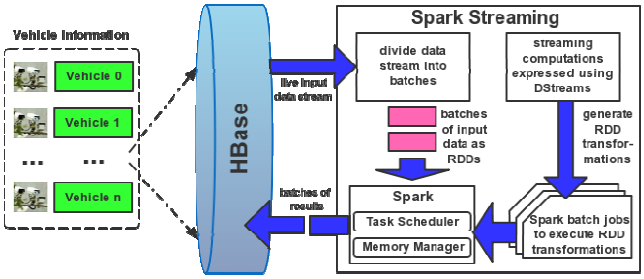


Figure 2. Our real time road saturation calculation framework

One the left of the diagram which represents to collect the vehicle license plate image. On the right is the process of sparking streaming which will include the following steps: we firstly divide data stream into batches as well as delete the valid data storage to the HBase database. Finally the road saturation real-time values are stored in HBase using the adjacent bayonet algorithm in Spark Streaming computing framework.

C. *The algorithm of adjacent traffic checkpoint*

1) Related definitions

a) Definition 1. *(The vehicle information collection)* The vehicle information sets will be collecting by the following formula.

$$E_{H_i}\{K_{i,j} \mid T_a \le T_{i,j} \le T_b, j = 1 \cdots M_i)\}(i = 1,2,\cdots,N) \quad (1)$$

Where E is the vehicle information collection, H_i is the car's license plate number i, N is the total number of vehicles, K is the number of traffic checkpoint, T is the vehicle through the traffic checkpoint time, T_a is the lower limit of time, T_b is a set time limit, $T_{i,j}$ is a time in setting time range, $K_{i,j}$ is the number of the car after the traffic checkpoint in time at $T_{i,j}$, M_i is the total number of the bayonet after the first i car.

b) Definition 2. *(Adjacent to the traffic checkpoint).* Suppose the two traffic checkpoint geographic distance less than 5 kilometers, then the two traffic checkpoint as a traffic checkpoint on the adjacent. The adjacent traffic checkpoint will be computed as follows.

$$\{K_n, K_m, C_B\} 1 < n < L, 1 < m < L, n \ne m \quad (2)$$

Where K_n and K_m represent the number of traffic checkpoint, C_B is on the road between the adjacent mount theoretical capacity, L is the total number of traffic checkpoint in road network.

2) The algorithm steps

The proposed algorithm of adjacent traffic checkpoint is to calculate the road saturation in the setting time of each road within the scope of the value by collecting vehicle information in the HBase database. The flow chart of the algorithm is shown in Fig.3, the main steps are as follows.

a) The value of each adjacent traffic checkpoint can be calculated as follows.

$$P(K_n, K_m) = \sum_{i=1}^{N} \sum_{j=1}^{M_{i-1}} q_{i,j} \qquad q_{i,j} = \begin{cases} 1, & K_{i,j}=K_n, K_{i,j+1}=K_m \\ 0, & other \end{cases} \quad (3)$$

b) The following equation can be used to compute f adjacent traffic checkpoint road traffic.

$$D(K_n, K_m) = \frac{P(K_n, K_m)}{T_b - T_a} \quad (4)$$

c) The adjacent traffic checkpoint at the set time range of road saturation can be written as

$$S(K_n, K_m) = \frac{D(K_n, K_m)}{C_B \gamma_l \gamma_r \gamma_c} \quad (5)$$

Where C_B is the theoretical capacity, γ_l is the width correction coefficient, γ_r is lateral clearance correction coefficient, γ_c is a heavy vehicle correction coefficient, specific values refer to the relevant specification manual.

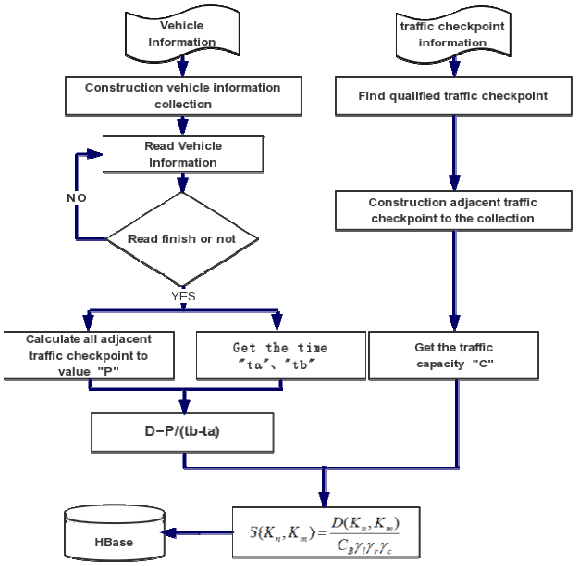


Figure 3. Flow chart of algorithm

Algorithm : The adjacent traffic checkpoint

Input: Vehicle information data, traffic checkpoint information data

Output: Road saturation set $S^{set}(K_n, K_m)$

Methods:

1. $E_{H_i} \leftarrow$ Build vehicle information collection
2. $\{K_n, K_m, C_B\} \leftarrow$ Build adjacent traffic checkpoint information
3. **for** each read Vehicle information collection in E_{H_i}
4. **for** each read adjacent traffic checkpoint information in $\{K_n, K_m, C_B\}$
5. **if** K_n, K_m conditions adjacent traffic checkpoint to meet
6. $P(K_n, K_m) \; += 1;$
7. **else**
8. Continue;
9. **if** After all read data vehicle information collection
10. **return** $P(K_n, K_m)$
11. **endif**
12. **endif**
13. **endfor**
14. **Endfor**
15. $D(K_n, K_m) = P(K_n, K_m)/(t_b - t_a)$
16. $S(K_n, K_m) = D(K_n, K_m)/C_B\gamma_l\gamma_r\gamma_c$
17. **return** $S^{set}(K_n, K_m)$

III. EXPERIMENTAL RESULTS

We chose a traffic checkpoint adjacent in Wenzhou city where the road saturation value will be updated at the interval of from 6 am to 5pm. The entire project is deployed in the Spark environment and the real-time calculation of road saturation by Spark under the framework of a sub frame Spark Streaming value, road saturation value is adjacent to the traffic checkpoint by the proposed algorithm. We have stored the data into non-relational database HBase which have been shown in Table 1, including the license plate number, after traffic checkpoint time, after traffic checkpoint number.

Table 1. The collection of vehicle information.

License plate number	After traffic checkpoint time	After traffic checkpoint number
浙 C0DX87	2015/2/1 15:47:00	3303040xx000
浙 CR8X39	2015/2/1 15:47:00	1010420xx000
浙 CRX100	2015/2/1 15:47:00	3303020xx000
浙 C2NX51	2015/2/1 15:49:00	3303040xx000
浙 C39X1D	2015/2/1 15:49:00	3303020xx000

We built a cluster of five nodes including a master node and four slave node. Traffic Flow-time is shown in Fig. 4 and the road saturation-time shown in Fig. 5.

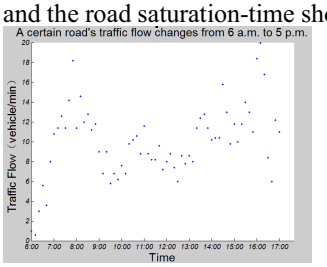


Figure 4. Traffic Flow-time

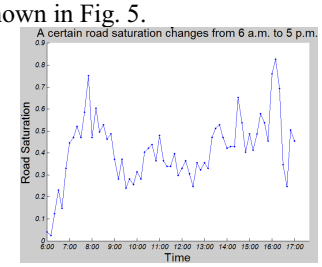


Figure 5. Road saturation-time

As shown in figure.6. We compare user satisfaction between our method with traditional method by the AB test for a period of 3 days (from April 22 2015 to April 24 2015), we randomly selected 500 scores of users. The test results of the first day is 423 people out of the calculation of the method in this paper is very satisfactory, compared with the traditional methods with 36 more people. And the test results of the second and third day of the number of satisfied method than the traditional method to more than 15.51% and 16.6% respectively.

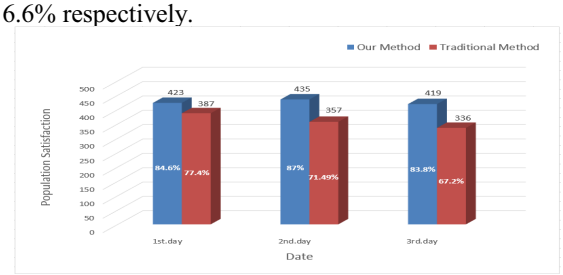


Figure 6. Comparison of this method with the traditional method of user satisfaction

IV. CONCLUSION

This paper presents a real-time calculation of road saturation method based on big data storage and computing that it is running in the big data processing engine Spark, while the source data and the result data is stored in a non-relational database in HBase. We will calculate the road traffic saturation data via adjacent traffic checkpoint while the computing and storage of road saturation based on Spark big data processing engine of big data and the final result data storage to the non-relational database HBase. The experiment results show that the performance of our method is superior to the original methods mainly embodied in the advantages of precision and real-time.

REFERENCES

[1] Marx. The big challenges of big data. Nature, vol. 498(7453), 2013, pp. 255-260.

[2] M. Minelli, M. Chambers, A. Dhiraj. Big Data, Big Analytics. Emerging Business Intelligence and Analytic Trends for Today's Businesses (Wiley CIO), 1st edition Wiley Publishing, 2013.

[3] Brown, B., Chui, M., Manyika, J.Are you ready for the era of 'big data'.McKinsey, vol. 4, 2011, pp. 24-35.

[4] Changjie Tang, Min Xiong, Changjie Tang, Min Xiong. The Temporal mechanisms in HBase. Journal of Computer Science and Technology, vol.11 (4), 1996, pp.365-371.

[5] Alejandro Vera-Baquero, Ricardo Colomo-Palacios, Owen Molloy. Towards a process to guide Big Data based Decision Support Systems for Business Processes. ProcediaTechnology, vol. 16, 2014, pp. 11-21.

[6] Isaac Triguero, Daniel Peralta, Jaume Bacardit, Salvador García, Francisco Herrera. MRPR: A MapReduce solution for prototype reduction in big data classification. Neurocomputing, vol. 150, 2015, pp. 331-345.

[7] Wang Peng, Qi Yan, Yang Hua-min. Based on the HBase query performance of database optimization research. Energy Education Science and Technology Part A: Energy Science and Research, vol. 32(4), 2014, pp. 2827-2834.

[8] Ashish Venugopal, Andreas Zollmann. Grammar based statistical MT on Hadoop: An end-to-end toolkit for large scale PSCFG based MT. The Prague Bulletin of Mathematical Linguistics, vol.91 (1), 2009, pp.67-78.

[9] Miguel A. Martínez-Prieto, Carlos E. Cuesta, Mario Arias, Javier D. Fernández. The Solid architecture for real-time management of big semantic data. Future Generation Computer Systems, vol. 47, 2015, pp. 67-29.

[10] Isaac Porche, Stephane Lafortune. Adaptive Look-ahead Optimization of Traffic Signals. Journal of Intelligent Transportation Systems, vol.4 (3), 1999, pp. 209-254.

[11] Nelson Sue. Big data: The Harvard computers. Nature, vol.455 (7209), 2008, pp.36-7.

[12] Zhou Guoliang, Zhu Yongli, Wang Guilan, Song Yaqi. Real-time big data processing technology application in the field of state monitoring. Diangong Jishu Xuebao Transactions of China Electrotechnical Society, vol. 29, 2014, pp. 432-437.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

An application of fuzzy rough sets in predicting on urban traffic congestion

Yingchao Shao
School of Information, Guizhou University of Finance and Economics, Guiyang, Guizhou, 550025, China
Email: shaoyingchao@sina.com

***Abstract*—In this paper, the soft fuzzy rough set theory is applied to predicting urban traffic congestion. For this purpose, a practical example predicting on urban traffic congestion based on the soft fuzzy rough set is presented.**

Key words–Rough set; Fuzzy rough set; Urban traffic congestion

I. Introduction

The rough set theory was initiated by Pawlak [7] for dealing with vagueness and granularity in information systems. This theory deals with the approximation of an arbitrary subset of a universe by two definable or observable subsets called lower and upper approximations. It has been successfully applied to machine learning, intelligent systems, inductive reasoning, pattern recognition, image processing, signal analysis, knowledge discovery, decision analysis, expert systems and many other fields [2], [8], [9], [10], [11].

In this paper we apply the notion of fuzzy rough set in predicting urban traffic congestion. This paper is organized as follows: Section 2 presents some basic concepts we use in this paper. Section 3 presents a method . Section 4 summarizes this paper.

II. Preliminaries

In this section, we first recall some fundamental facts about Pawlak's rough sets[7].

***Definition** 1:* [9] An information system is a pair $\mathcal{I} = (U, R)$ of non-empty finite sets U and A, where U is a set of objects and A is a set of attributes; each attribute $a \in A$ is a function $a : U \to V_a$, where V_a is the set of values(called domain) of attribute a.

Let U be a non-empty finite universe and R be an equivalence relation on U. The pair (U, A) is called a Pawlak approximation space. The equivalence relation R is often called an indiscernibility relation and related to an information system. Specifically, if $\mathcal{I} = (U, A)$ is an information system and $B \subseteq A$, then an indiscernibility relation $R = I(B)$ can be defined by

$$(x; y) \in I(B) \Leftrightarrow a(x) = a(y); \forall a \in B;$$

where $x, y \in U$, and $a(x)$ denotes the value of attribute a for object x.

Using the indiscernibility relation R, one can define the following two operations

$$R_{-}(X) = \{x \in U | [x]_R \subseteq X\}; R^{-}(X) = \{x \in U | [x]_R \cap X \neq \emptyset\}$$

assigning to every subset $X \subseteq U$ two sets $R_{-}(X)$ and $R^{-}(X)$ called the R-lower approximation of X and the R-upper approximation of X, respectively.

If $R^{-}(X) = R_X$, then X is called a definable set; if $R_{-}(X) \neq R^{-}(X)$, then X is called an undefinable set, and (R_X, R^X) is referred to as a pair of rough set, or a rough set.

III. A Method predicting on urban traffic congestion based on the soft fuzzy rough set

In this section, we will give an application of fuzzy rough sets in predicting on urban traffic congestion.

A. The basic ideal

It is well known that there are many factors which causes the urban traffic congestion such as road, weather, time, unforeseen event,etc. Based on [3], the factor of road is redundancy, that is, its effect to traffic congestion can be reflected by the other factors, so we can ignore it in the research. The effect caused from different factors is different, so we can choose some important factors to predict the urban traffic congestion.

For this purpose, we set up a set of the factors, denoted by U. Because of these factors are interconnecting, we can express the relation by a fuzzy relation R. So we can set up a fuzzy approximation space (U, R). And then we set up a fuzzy rough set to indicate the congestion. By consulting the observer, we can determine the traffic conditions of this road in the past, and then set up a mapping from the traffic condition to a fuzzy rough set, that is, we set a soft fuzzy rough set. We indicate the traffic condition of this road by a fuzzy rough set. At last, we can predict the traffic congestion by a matching function.

B. The predictive steps

1) Set a traffic data table.
   The table shows the congestion record of one road in past, including various factors effecting the traffic condition, such that time, emergency, weather, entryman, etc.
2) According to the expert opinions, transform the traffic data table into a figure table.
   In general speaking, the traffic record data table is expressed with the natural language, for example, the time factor is expressed with morning-evening rush hours, holiday rush etc, which requests they must be

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.59

transformed into figures so that they can be computed. It can be done by mean of some expert opinions.

3) find the interrelation of various factors and set a fuzzy approximation space (U, R), where, U denote a set of various factors, and R denotes a fuzzy relations between the factors. In this paper the set of the factors is set based on the paper [3] as following:

$$U = \{time, weather, entryman, unforeseen \quad event\}.$$

Many researchers presented many different methods getting the fuzzy relation R. We present a method getting the R based on the papers [2], [14], [1] as following:

$H(t) = \sum_{i=1}^{n} \mu_i(t) \log \mu_i(t)$,

$$\mu_{i,j}(t,s) = \begin{cases} max\{\mu_i(t), \mu_j(s)\}, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

$H(t,s) = -\sum_{i,j} \mu_{i,j}(t,s) lg \mu_{i,j}(t,s)$,

$I(t,s) = H(t) + H(s) - H(t,s)$,

where, $t \in U, S \in U, \mu_i$ denotes the observed value the factor t effects the traffic congestion at the n time. Suppose that any two factors is independent in the different observation. $\mu_{i,j}(t,s)$ denotes the effect degree that the factor t and the factor s effect the traffic congestion in the observation at the i time and the j time, respectively. $H(t,s)$ denotes a expectation that the factor t and the factor s effect the traffic congestion in the n times observations. $I(t,s)$ is a mutual information. It reports the interrelation between the factor t and the factor s. Its value is more big, it shows the effect incident to the congestion is more adjacent.

$I(t,s)$ has the following properties:

(i)nonnegative, i.e., $I(t,s) \geq 0$;

(ii)symmetrical, i.e., $I(t,s) = I(s,t)$;

(iii) $I(t,t) = H(t)$.

In order to compare the dependency of two factors, we use the following formula presented in [1] to express the relation between two factors t and s:

$$R_{t,s} = \frac{I(t,s)}{\sqrt{H(t)H(s)}}, \tag{1}$$

It is clever that $R_g(t,t) = 1, 0 \leq R_g(t,s) \leq 1$.

4) Let a fuzzy rough set $(\underline{Q_{I_i}}, \overline{Q_{I_i}})$ express the traffic congestion.

Firstly, express the threshold set of congestion with a fuzzy set by consulting the experts as following:

$I_i = (\omega_{i1}/t_1, \omega_{i2}/t_2, \cdots, \omega_{in}/t_n)$,

these congestions may be open, mild jams, ordinary jams, congestion, that is, $I_i \in \{open, mildjams, ordinaryjams, congestion\}$. And then get the fuzzy rough set $(\underline{Q_{I_i}}, \overline{Q_{I_i}})$ by the following formula:

$\underline{Q_{I_i}(t)} = \bigwedge_{s \in U} [(1 - R(s,t)) \vee I_i(s)], Q_{I_i}$

$= \bigvee_{s \in U} [R(t,s) \wedge I_i(s)]$,

where, $I_i(s)$ denotes the degree of membership of the factor s in the fuzzy set I_i. For example, $I_i(t_1) = \omega_{i1}$; $\underline{Q_{I_i}(t)}$ denotes the minimal value that the factor t effects I_i; $\overline{QI_i(t)}$ denotes the maximum value that the factor t effects I_i.

Let $(\underline{Q_{I_i}}(t), \overline{QI_i}(t))$

$= ((\underline{Q_{I_i}}(t_1), \overline{QI_i}(t_1)), (\underline{Q_{I_i}}(t_2), \overline{QI_i}(t_2)),$

$\cdots, (\underline{Q_{I_i}}(t_n), \overline{QI_i}(t_n)))$

***Remark** 1:* The threshold set Ii denotes a expectancy various factors effect the traffic congestion. It can be gotten by consulting the related experts, or by calculating arithmetic average according to historical record on the road.

5) Set a soft fuzzy rough set.

In general speaking, it is in the natural language that an observer describes the situation of a road section. For example, usual time interval, emergency, foggy, etc. In fact, he gives an element x of the set E, and it is required a mapping transforming the natural language into a fuzzy value of U:

$f : E \to F(U)$,

$f(x) = \mu(x)$,

where, E denotes a language vector set, $x \in E$, $\mu(x)$ denotes the fuzzy vector set on U. For example, Let $E = \{(t_1, t_2, t_3, t_4) j t_1 \in E_1, t_2 \in E_2, t_3 \in E_3, t_4 \in E_4\}, \mu(x) = (\mu(t_1, \mu(t_2, \mu(t_3, \mu(t_4,).$

where, E_1 denotes a set of the factor time in different time interval, for example, $E_1 = \{off - peak, rush\ hours, peakholiday\}$;

E_2 denotes a set of different kinds of emergency, for example,

$E_2 = \{normal, road\ work, traffic$

$accident, trafficcontrol\}$,

E_3 denotes a set of weather conditions, for example,

$E_3 = \{sunny, rain, foggy, snow\}$

E_4 denotes a set of entryman forms, for example,

$E_4 = \{normal, illegalparking, shop'spresence\}$.

By step (5), determine its fuzzy rough set $(\underline{q}(t_n), \overline{q}(t_n))$:

$(\underline{q}(t_i), \overline{q}(t_n)) = \bigwedge_{s \in U} [(1 - R(t_i, s)) \vee \mu(s)]$;

$(\overline{q}(t_i), \overline{q}(t_n)) = \bigvee_{s \in U} [R(t_i, s) \wedge \mu(s)]$.

So we set a soft fuzzy rough set. By the soft fuzzy rough set, we can get a fuzzy rough set of traffic congestion observe value x at some time.

6) Calculate the relevancy of $(\underline{Q}(t), \overline{Q}(t))$ and $(\underline{q}(t), \overline{q}(t))$ by a matching function, and then according to the value, judge if the traffic congestion happens at the road.

We choose the matching function defined in [14]:

$$SIM(Q,q) = \underline{SIM}(Q,q) + \overline{SIM}(Q,q), \tag{2}$$

where,

$$\underline{SIM}(Q,q)=\frac{\sum_{i=1}^{n}(\underline{Q_{I_i}}(t))\wedge\underline{q_i}(t)}{\sum_{i=1}^{n}(\underline{Q_{I_i}}(t))\vee\underline{q_i}(t)},$$

$$\overline{SIM}(Q,q)=\frac{\sum_{i=1}^{n}(\overline{Q_{I_i}}(t))\wedge\overline{q_i}(t)}{\sum_{i=1}^{n}(\overline{Q_{I_i}}(t))\vee\overline{q_i}(t)}.$$

IV. ANALYSIS OF EXAMPLES

According to [16], the factors effecting congestion mainly are time(Off-peak times, rust hours, peak holiday), emergency(road work,traffic accident, traffic control),whether(sunny, snow, foggy, rain, wind), entryman(normal, illegal parking, shop's presence). So we can set a factor set
$U=\{time, emergency, wheather, entryman\}$.
we predict the traffic congestion of a road section present in [16] in the following: Set membership tables II,III,IV,V

TABLE I
THE TRAFFIC INFORMATION TABLE

	time	*emergency*	*wheather*	*entryman*	*condition*
1	*offpeak*	*normal*	*sunny*	*normal*	*open*
2	*offpeak*	*road work*	*rain*	*illegal parking*	*mild jams*
3	*rushhours*	*accident*	*foggy*	*normal*	*congestion*
4	*rushhours*	*normal*	*foggy*	*normal*	*mild jams*
5	*off − peak*	*normal*	*rain*	*normal*	*open*
6	*rush hours*	*normal*	*sunny*	*normal*	*mild jams*
7	*holiday peak*	*control*	*sunny*	*illegal parking*	*congestion*
8	*holiday peak*	*control*	*rain*	*shop′spresence*	*congestion*
9	*offpeak*	*normal*	*rain*	*illegal parking*	*mild jams*
10	*offpeak*	*road work*	*snow*	*illegal parking*	*congestion*

according to the experts' opinions[17]:

TABLE II
THE MEMBERSHIP TABLE OF THE FACTOR TIME

time	*off − peak*	*rush peak*	*holiday peak*
membership	0.3	0.6	0.8

TABLE III
THE MEMBERSHIP TABLE OF THE FACTOR WEATHER

weather	*snow*	*storm*	*foggy*	*rain*	*sunny*
membership	1.0	0.9	0.5	0.7	0

TABLE IV
THE MEMBERSHIP TABLE OF THE FACTOR ENTRYMAN

entryman	*illegalparking*	*shop′spresence*	*normal*
membership	0.4	0.6	0

So we can transform the table I into the following figure table VI:

So we can get the fuzzy relation R on U by the formula 1 as follow:

The threshold sets which reflects congestion occurs in the road section L are
$L_{congestion}=\{0.625/time, 0.475/emergency,$
$0.275/weather, 0.35/entryman\}$;
$L_{mildcongestion}=\{0.5/time, 0.3/emergency,$

TABLE V
THE MEMBERSHIP TABLE OF THE FACTOR EMERGENCY

emergency	*road work*	*accident*	*control*	*normal*
membership	0.7	0.6	0.5	0

TABLE VI
THE TRAFFIC INFORMATION TABLE

	time	*emergency*	*wheather*	*entryman*	*condition*
1	0.3	0	0	0	*open*
2	0.3	0.7	0.5	0.4	*mild jams*
3	0.6	0.6	0.6	0	*congestion*
4	0.6	0	0.6	0	*mild jams*
5	0.3	0	0.5	0	*open*
6	0.6	0	0	0	*mild jams*
7	0.8	0.5	0	0.4	*congestion*
8	0.8	0.5	0.5	0.6	*congestion*
9	0.3	0	0.5	0.4	*mild jams*
10	0.3	0.7	1	0.4	*congestion*

TABLE VII

L	time	emergency	weather	entryman
time	1	0.801	0.971	0.751
emergency	0.801	1	0.721	0.887
weather	0.971	0.721	1	0.768
entryman	0.751	0.887	0.768	1

$0.45/weather, 0.3/entryman\}$;
L_{open}
$=\{0.3/time, 0.1/emergency, 0.3/weather, 0.2/entryman\}$.

Then the threshold fuzzy rough set which reflects the congestion occurs in the road L is in the table VIII

TABLE VIII

L	Time	Emergency	Weather	Entryman
$(\underline{Q_L},\overline{Q_L})$	(0.275,0.625)	(0.279,0.625)	(0.279,0.625)	(0.279,0.625)

The threshold fuzzy rough set reflects the mild jams occurs in the road L is in the table IX:

TABLE IX

L	Time	Emergency	Weather	Entryman
$(\underline{Q_L},\overline{Q_L})$	(0.2,0.45)	(0.2,0.45)	(0.232,0.45)	(0.2,0.45)

The threshold fuzzy rough set which reflects the open occurs in the road L is in the table X:

TABLE X

L	Time	Emergency	Weather	Entryman
$(\underline{Q_L}(t),\overline{Q_L}(t))$	(0.199,0.3)	(0.1,0.3)	(0.232,0.3)	(0.113,0.3)

***Example** 1:* Suppose an observer describes the conditions of the road L as "rush hours, no emergency, rain, shop's presence ". In fact, he represents a parameter $e = (rush \quad hours, normal, rain, shop'spresence)$. According to this, we can set a mapping f such that $f(e) = \{0.9/time, 0/emergency, 0.4/weather, 0.7/entryman\}$ so we can get the fuzzy rough set q_L on L at the time as follow:

TABLE XI

L	Time	Emergency	Weather	Entryman
$(\underline{q}_L(t),\overline{q}_L(t))$	(0.199,0.9)	(0,0.801)	(0.279,0.9)	(0.113,0.751)

By the formula 2, we can get that
$SIM_{mildlams}(Q_L, q_L)$
$= \underline{SIM}_{mildlams}(Q_L, q_L) + \overline{SIM}_{mildlams}(Q_L, q_L)$
$= \frac{\sum_{i=1}^{n}(\underline{Q}_L(t))\wedge\underline{q}_L(t))}{\sum_{i=1}^{n}(\underline{Q}_L(t)\vee\underline{q}_L(t))} + \frac{\sum_{i=1}^{n}(\overline{Q}_L(t)\wedge\overline{q}_L(t)}{\sum_{i=1}^{n}(\overline{Q}_L(t)\vee\overline{q}_L}(t))$
≈ 1.089;
$\underline{SIM}_{congestion}(Q_L, q_L) = \underline{SIM}_{congestion}(Q_L, q_L) + \overline{SIM}_{congestion}(Q_L, q_L) \approx 1.276$;
$\underline{SIM}_{0pen}(Q_L, q_L) = \underline{SIM}_{open}(Q_L, q_L) + \overline{SIM}_{open}(Q_L, q_L) \approx 1.145$;
So we can predict that congestion will occur in the road L.

V. CONCLUSION

In this paper, an application of fuzzy rough set in predicting urban traffic congestion is presented. It is worth mentioning that the method presented in this paper has some defects, such as, the fuzzy relation R derived by this way is of subjective and the algorithm is too complex. How to overcome these defect will be an intent we further study.

ACKNOWLEDGMENT

This work was supported by 2013 the Doctor funds of Guizhou University of Finance and Economics.

REFERENCES

[1] J.Ding, W.wang, Y.Zhao, *General Correlation Coefficient Between Variables Based on Mutual Information*, Journal of Sichuan University(Engineering Science Edition),2002,34(3):1-5.
[2] X.Fu, Q.Liu, M.Wang, *Rough sets information retrieval model based on mutual information*, Journal of Shandong University. 2006,41(3):17-19.(In Chinese).
[3] L.Fang, L.Wei, W.Yan, *Prediction of Traffic Congestion Based on Decidition Tree*, Journal of Hebei University of Technology, 2010,39(2):105-110.
[4] P. Hájek , *Metamathematics of Fuzzy Logic*, Kluwer Academic Publishers. 1998.
[5] L.Li, X.Zhang,*Research on a new type of information retrieval system based on rough set*, Journal of The China Society For Scientific and Technical Information, 2002,21(1):7-11.
[6] X.Ma, J.Zhan, *On($\in, \in \vee q$)-fuzzy Filters of Bl-algebras*, Jrl Syst Sci & Complexity. (2008)21:144-158.
[7] Z.Pawlak, *Rough sets*, Int.J.Comput. Inform.Sci.11(1982)341-356.
[8] Z.Pawlak, *Rough sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers., Dordrecht, 1991.
[9] Z.Pawlak, A.Skowron, *Rudiments of rough sets*, Inform. Sci. 177(2007)3-27.
[10] Z.Pawlak, A.Skowron, *Rough sets: some extensions*, Inform. Sci. 177(2007)28-40.
[11] Z.Pawlak, A.Skowron, *Rough sets and Boolean reasoning*, Inform. Sci.177(2007)41-73.
[12] P.Srinivasan, *The importance of rough approximations for information retrieval[J]*. International Journal Man-Manchine Studies,1991,22(34):657-671.
[13] P.Srinivasan, *Intelligent information retrieval using rough set approximations[J]*, Information Processing and Management, 1989,25(4):347-361.
[14] D.Tan, *Research on Fuzzy Rough Sets for Application in Technology Document Concept Retrieval*, Computer Simulation, 2011,28(10):168-172.
[15] D.Wu, H.Qi, *Development of Information Retrieval Model and its Application in Cross-Language Information Retrieval*, Journal of Modern Information. 2009, Vol.29, No.7, 215-221.
[16] Y.Zhang, P.Lu,*Algorithm Research of Urban Traffic Congestion Early Warning Based on Rough Set Theory*, Technology and Economy in Areas of Communications, 2009,2:74-76.
[17] Y.Zhang, C.Yi, P.Wang, *The discrimination Method Used for Traffic Jam Based on Fuzzy Inference*,Traffic Engineering, 2012,02:64-69.

Network Performance of EPA Protocol Based on Simulation Tool

Ping Zhou[1]

[1]Wuhan Nanhua Industrial Equipments Engineering CO.,LTD,

Wuhan, China

e-mail: 981550980@qq.com

Yuqing Feng[2]

[2]Wuhan Social Work Professional College, Wuhan, China

e-mail: 280064139@qq.com

Abstract: EPA (Ethernet for Plant Automation) resolves the nondeterministic problem of Ethernet through sub-segment topology and deterministic scheduling mechanism. In this paper, the EPA scheduling mechanism is implemented by simulation tool. And the influence of EPA scheduling mechanism on network performance is shown in the simulation result. According to the study, the real-time performance of EPA periodic messages transmission is concluded and the transmission delay of non-periodic message is analyzed.

Keywords: EPA, simulation, scheduling mechanism, network performance

1. INTRODUCTION

EPA(Ethernet for Plant Automation) real-time Ethernet is a new kind of open real-time Ethernet standards, used in industrial field devices. It has been published by IEC as IEC/PAS 62409 and specified in IEC 61784-2 as the communication profile family 14 (CPF 14) and will be integrated into the revised edition 4 of IEC 61158. EPA applies a lot of mature IT technologies to the industrial control system, uses effective, stable, standard Ethernet and determinable communication scheduling strategy of UDP/IP protocols and sets a new standard for the real time work of the field devices.

In industrial field, there are different requirements for different applications. Performance indicators shall be used to specify capabilities of real-time communication network as well as to specify requirements of an application. Performance indicators will be used as a set of interaction means for evaluating the network design and architecture. There are 9 PIs defined in IEC 61784-2: Delivery time, Number of end nodes, Basic network topology, Number of switches between end nodes, Throughput RTE(Real-Time Ethernet), Non-RTE bandwidth, Time synchronization accuracy, Non time-based synchronization accuracy, Redundancy recovery time, as in [4]. However, the analysis on the real-time performance of EPA has just begun in theory and simulation methods. It is necessary for the performance of EPA's scheduling mechanism to be analyzed by simulation way so that EPA could meet the application requirement of industry and get further improvement.

OPNET is one of the most popular tools due to its powerful functions and excellent customer support [3]. The Modeler, one of the many tools in the OPNET Technologies suite, whose engine is based on a finite state machine model, can model protocols, devices, and behaviors with abundant purpose functions. And it provides a number of editors to simplify different levels of modeling essential for the network operators. Nowadays the OPNET Modeler is widely used for protocol network performance[5].

2. EPA SCHEDULING MECHANISM

Ethernet for Pant Automation (EPA) is a new real-time Industrial Ethernet-based network. EPA fully complies with the IEEE802.3 Ethernet standard. In order to improve the real-time property of communication system, a real-time communication schedule interface is added between the Network Layer and the Data Link Layer, CSME (communication schedule management entity) is used to cope with the uncertainness aroused by the CSMA/CD mechanism of Ethernet, so it can provide data communication that is confirmed and real-timed for the Industrial Ethernet. A simplified IEEE1588 Standard (Precision Clock Synchronization Protocol) is used to keep all devices drives uniform on time. So in the communication schedule interface, the time to complete a communication procedure is called communication macrocycle. In Fig. 1, each communication macrocycle (T) is divided into two phases, periodic message transferring phase (Tp) and non-periodic message transferring phase (Tn). During Tp, the time-critical payload, usually it is function block output data, is sent in its own slot (this slot is an offset time relative to the beginning of macrocycle). During Tn, non-time-critical payload, such as service data for reading response, writing response, notification information and device configuration, is sent orderly by its priority and IP address. The last part of each device's periodic message contains a non-periodic message announcement for begin which indicates whether the device has a non-periodic message to transmit or not. Once the non-periodic message transferring phase begins, all devices which have announced (during the periodic message transfer phase) that they had a non-periodic message to send are allowed to transmit their non-periodic messages. If one device finishes transmitting, it must transmit a non-periodic message announcement for end to indicate other devices.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.60

3. SIMULATION MODEL

According to the above EPA protocol analysis, corresponding network modules, node modules and Process Modules based EPA protocol are built in OPNET.

A. Node Model

Because we just want to get the network performance, the EPA node model is simplified to only include the lower layers below the IP layer, shown in Fig. 2.

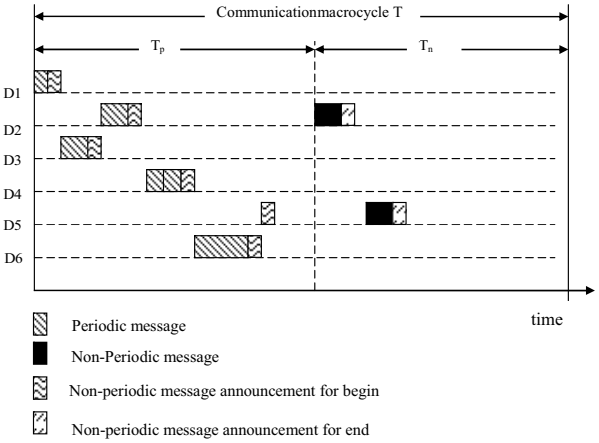


Fig. 1 Time division of macrocycle and scheduling order of EPA

Most of the Ethernet Modules are existed in OPNET Model, such as the mac module and defer module realized CSMA/CD protocol. simp_const and simp_random model can generate periodic and non-periodic data (according to Possion distribution)，respectively. queue_fifo is an first-in first-out queue, and queue_prio is a prioritized queue. epa_schedule module implements the EPA schedule mechanism, and eth_mac_intf provides the Clock Synchronization. The others modules are simulated the Ethernet CSMA/CD medium access mechanism. In Fig. 2, physical layer is bus topology, and indeed the others Ethernet topology has been realized by OPNET Model, we are simply instead of physical layer model to get them.

B. Process Module

The main task of Process Module of EPA schedule is to realize periodic and non-periodic task dispatch. The time is divided into two parts, periodic state takes charge for periodic tasks schedule, nonperiodic state takes charge for non-periodic tasks schedule. In the phase of periodic task, when the sending time of node reaches, request_0 state requests data from queue_fifo, and send_0 state sends data to eth_mac_intf module, then enter the endpkt state at the end of sending periodic data, to send the priority data field, announcing the whole network the current priority of this node, making preparation for non-periodic tasks. The frommac state receives data from the eth_mac_intf module. decide state is used to decide if the duration of Tp or Tn have been finishing. The process of non-periodic tasks is similar to periodic tasks. The difference is after entering the state of send, it changes into the priority state instantly, sending the priority data field. Only the node with the highest priority can send data, so collisions can be avoided effectively.

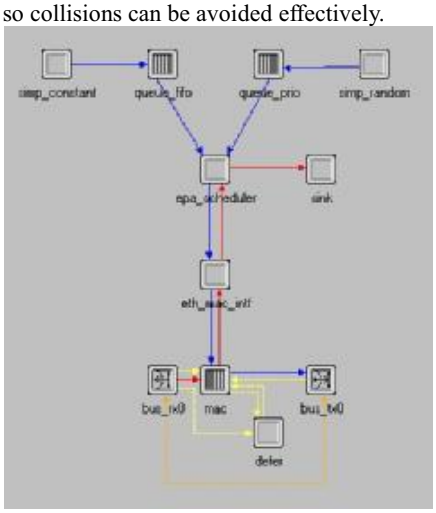


Fig. 2 node model

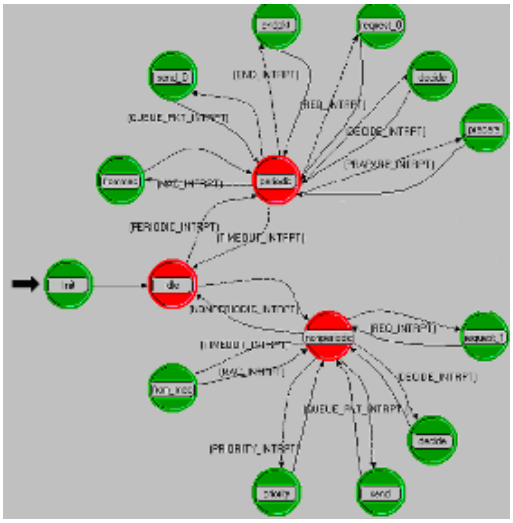


Fig. 3 EPA_scheduler process module

4. PERFORMANCE EVALUATION

A. Network Performance Index

For evaluating the network performance, those performance indexes such as Data Successful Transmission Percent, Data Transmission Delay Time, can be used.

1) Data Successful Transmission Percent: Due to EPA protocol is still based on CSMA/CD, whether the collision can be avoided by EPA scheduling mechanism is evaluated by this index.

2) Data Transmission Delay Time: The duration is from the beginning time of data generation to the ending time of data received by the other nodes. Using

this index, we can estimate if the periodic data can be received in the predefined time.

B. Simulation Result

Using the established network simulation platform, we present the simulation scenario as showing in the above. The simulation results are shown in Fig. 4 and 5.

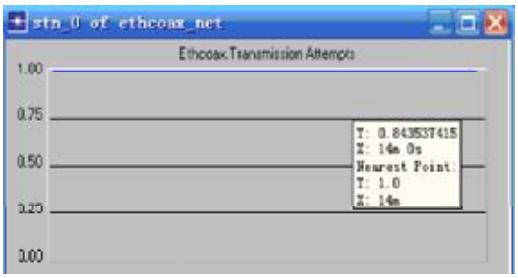


Fig. 4 simulation results of Data Sucessful Transmission Percent

In Fig. 4, all packets of node 0, even the non-periodic packet can be once sent successfully. It proves that the EPA scheduling mechanism is effective to avoid the collisions on the data link layer of CSMA/CD.

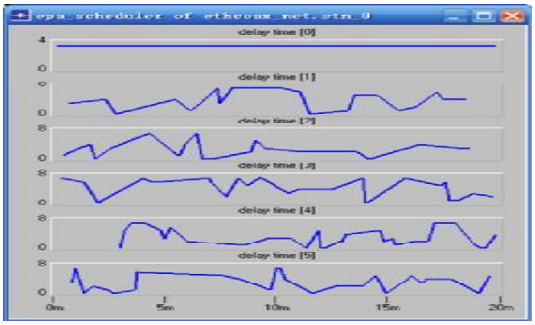


Fig. 5 simulation results of Data Transmission Time

In Fig. 5, it shows the transmission delay time of different priority packet in node 0. delay_time[0] representing periodic data with priority 0, is a constant value. It shows that for periodic data, the transmission delay is determinate time, and it is suitable to critical transmission requirements. On the contrary, the non-periodic packet cannot be sent by priority order and delay times are variable, because in non-periodic phase, the highest priority perhaps is not appeared at the moment. Node can know other nodes' non-periodic data information by receiving the packet from other nodes. In some case, highest non-periodic data cannot be sent in time, and has a bigger delay. In Fig. 6, an example of EPA non-periodic packet transmission is shown to illustrate the above case.

Supposing that there are two nodes D1 and D2 for transmitting message, and they generated several non-periodic messages with different priorities in Tp phase. The non-periodic message transmitting order is sorted by the information from non-periodic message announcement for begin. So in Fig. 6, at the beginning of Tn, the node D2 only got the information that the node D1 has a message with priority 4 to send. And it considers that it owns the higher priority message and sends firstly. Although node D1 has the message with priority 1 to send, this message generated after the message of non-periodic message announcement for begin. So this information cannot be known by other nodes. Node D1 follows the order as the case without this message with priority 1. But in its turn to send the message, node D1 will send the message with priority 1 instead of priority 4. Even the non-periodic message transmission cannot be in accordance with priority; this mechanism is still suitable for non-real-time communication. And because the collision is avoided, the phase time can be guaranteed.

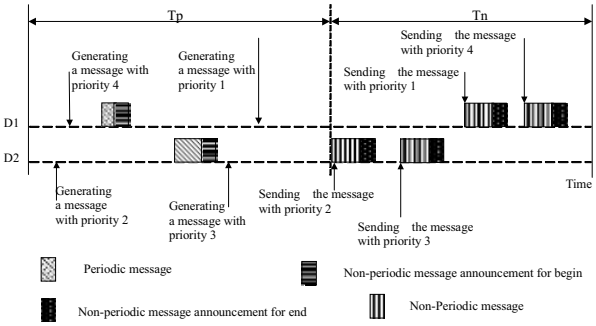


Fig. 6 non-periodic packet transimission processdure

5. CONCLUSION

In this paper, the network performance of EPA scheduling mechanism is evaluated using OPNET simulation methods. It proves that the collision can be avoided, and the periodic transmission has determinate transmission time, and non-periodic transmission has the variable transmission time and cannot be transmitted in accordance with priority.

References

[1] EPA system architecture and communication specification for industrial control and measurement systems, General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, GB/T 20171-2006, 2006.

[2] Liu Ning, Zhong Chongquan, Analysis on Scheduling Performance of Periodic Messages in Industrial Ethernet Based on EPA, Proceedings ofthe 2009 IEEE International Conference on Mechatronics and Automation August 9-12, Changchun, China

[3] M. Cheng, OPNET network simulation, Beijing: Tsinghua University Press, 2004

[4] Ludwig Winkel, Siemens AG, Karlsruhe, Germany, Ludwig, Real-Time Ethernet in IEC 61784-2 and IEC 61158 series, 2006 IEEE International Conference on Industrial Informatics, pp. 246-250.

[5] Zhang Shuai, Feng DongQin, Chu Jian. Probability-based clock synchronization for EPA Wireless protocol. Journal of Zhejiang University (Engineering Science), 2014,48(9):1552-1557.

A Distributed Power-Saving Topology Management Scheme in Green Internet

Jinhong Zhang, Xingwei Wang, Min Huang
College of Information Science and Engineering
Northeastern University
Shenyang, 110819, P.R. China
neuzjh@aliyun.com; wangxw@mail.neu.edu.cn; mhuang@mail.neu.edu.cn

***Abstract*—In a context of explosive development in Internet scale and excessive underutilization of a large number of network elements, the power consumption issues of the current Internet have become hot subjects extensively concerned by ICT (Information and Communication Technology) sector and research institutions in recent years, which even further lead to the severe economic issues and intractable environmental issues. On the basis of above circumstances, a Distributed Power-saving Topology Management Scheme (DPTMS) is devised to achieve power saving for current Internet in the paper. After introducing the network model and power model, we describe the design of DPTMS including four modules: traffic prediction module, information awareness model, distributed topology decision-making module and sleeping control module. In the simulation and evaluation, DPTMS is compared with GRiDA as for network power consumption and network performance, which shows it is feasible and effective for a power-saving topology management.**

Keywords-green Internet; power saving; distributed decision-making; topology management; VCG mechanism

I. INTRODUCTION

Due to the rapid development of Internet scale, the power consumption in Internet has an amazing growth in the recent years. The power consumption in Internet will increase 50% in USA for the upcoming ten years and account for 8% of the total power consumption in USA [1].

From a network-wide point of view, studying network topology management is a way to achieve the power saving in current Internet. Ref. [2] proposed two forms of topology management schemes which are putting network components to sleep during idle times and adapting the rate of network operation to the offered workload to reduce the energy consumption of networks. Ref. [3] proposed a three-phase algorithm in the topology management: in the first phase some routers are elected as exporter of their own Shortest Path Trees; in the second one the neighbors of these routers perform a modified Dijkstra algorithm to detect links to power off; in the last one new network paths on a modified network topology are computed. Ref. [4] provided two new algorithms for exploiting low traffic load patterns commonly found in Ethernet switches and endpoints, and opportunistically powered down unused network interfaces in order to save some of that wasted energy. Ref. [5] formulated power-saving topology management as a capacitated multi-commodity flow problem and proposed some heuristic algorithms to switch off network nodes and links while still guaranteeing full connectivity and maximum link utilization. Ref. [6] proposed and assessed strategies to concentrate network traffic on a minimal subset of network resources, aiming to turn off network nodes and links while still guaranteeing full connectivity and maximum link utilization constraints. Ref. [7] presented generic models and numerical results that indicate dynamic topologies are able to reduce power consumption of communication networks by adapting topologies to traffic conditions. Ref. [8] proposed a comprehensive approach to determine a network topology and a link metric for each time period and a distributed loop-free routing update scheme to determine the correct sequence for updating the routing table, and solved a multi-topology and link weight assignment problem.

Based on the present state of study, a distributed power-saving topology management scheme-DPTMS is proposed.

II. PROBLEM DESCRIPTION

A. Network Model

The network model is described as a simplified connected graph $G(V, E)$, where V represents a set of vertexes (nodes) and E represents a set of edges (links). The node structure is depicted in Fig. 1, consisting of chassis, line cards, a master engine (ME), forwarding engines (FE), replication engines (RE) and so on [9]. The link structure is depicted in Fig. 2, consisting of a pre-amplifier, in-line amplifiers, regenerators, a post-amplifier and so on [10].

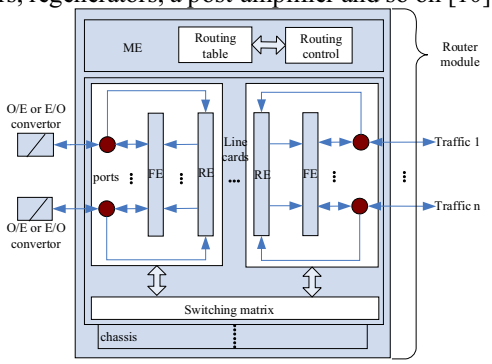


Figure 1. Node structure

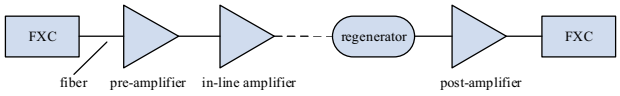


Figure 2. Link structure

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.61

B. Power Model

The power models of a node and a link are shown in (1) and (2) respectively, of which the meanings of the notations involved are stated in Table I and Table II.

TABLE I. NODE PARAMETERS

Notations	Meanings
$P^i_{n\_ctrl}$	the power consumed by the master engine in the core router i
$P^i_{n\_frm}$	the power consumed by a chassis in the core router i
$P^l_{l\_cpu}$	the power consumed by CPU of the line card in the core router i
$P^l_{l\_mem}$	the power consumed by memory of the line card in the core router i
$P^l_{l\_bus}$	the power consumed by bus of the line card in the core router i
P^p_{port}	the power consumed by a port in the core router i
N^i_{from}	the number of chassis in the core router i
N^k_{lcrd}	the number of line cards of chassis k in the core router i
N^l_{port}	the number of ports in line card l of chassis k in the core router i
trf_l	the traffic flowing on line card l in the core router i
trf_p	the traffic flowing on port p in the core router i
α, β	the constants used to confirm the relation between traffic and power consumption

TABLE II. LINK PARAMETERS

Notations	Meanings
$lkpw_j$	the power consumption of link j
len_{ref}	the referential length of link (the default is 80km)
len_j	the practical length of link j
P_{rely}	the benchmark power consumption
trf_j	the traffic in link j
α, β	the constants used to confirm the relation between traffic and power consumption

$$ndpw_i = P^i_{n\_ctrl} + \sum_{k=1}^{N^i_{frm}} \left(P^i_{n\_frm} \times FrmSt^k_i + \sum_{l=1}^{N^k_{lcrd}} \left(2 \times \left(P^l_{l\_mem} + P^l_{l\_cpu} \times \left(1 + \alpha \times trf_l^{\beta}\right)\right) \times LdSt^l_k + \sum_{p=1}^{N^l_{port}} \left(P^p_{port} \times (1 + \alpha \times trf_p^{\beta}) \times PortSt^p_l \right) \right) \right) \quad (1)$$

$$lkpw_j = \left\lfloor len_j / len_{ref} \right\rfloor \times P_{rely} \times \left(1 + \alpha \times trf_j^{\beta}\right) \quad (2)$$

III. SCHEME DESIGN

A. Traffic Prediction Module

The traffic prediction module aims to predict the port traffic by the prediction algorithm on the basis of the history traffic data recorded in the port.

The traffic prediction procedure is shown as follows:

Step1：Check the port of node, and if the time arrives a threshold $tmth$, execute traffic prediction; else quit.

Step2：Further predict by applying the autoregression method and the Markov method. If both of them are invalid, then go to Step1; if the only one of them is valid, then take its value as the prediction result; if both of them are valid, then take their average as the prediction result.

Step3：Is it executed N time? If no, go to Step2.

Step4：Discard the top 5% of results in N results, and take the maximum in the remaining results as the final prediction result.

B. Information Awareness Module

The information awareness module is in charge of managing the node awareness and traffic awareness.

The change extent $tfch$ of traffic is calculated in (3), where trf_{last} represents the last traffic, trf_{now} the current traffic, $0<\theta<1$. The current awareness interval $tfin$ is calculated in (4), where $tfin_{last}$ represents the last one, $tfin_{st}$ represents the stable one, $0<\theta<1$, γ is an increasing coefficient; $pmin$ represents the ratio of minimum one to stable one.

$$tfch = \begin{cases} \left| trf_{last} - trf_{now} \right| / trf_{last}, & trf_{last} \neq 0 \\ 0, & trf_{last} = 0, trf_{last} = trf_{now} \\ \theta, & trf_{last} = 0, trf_{last} \neq trf_{now} \end{cases} \quad (3)$$

$$tfin = \begin{cases} \theta \times tfin_{last}, & tfch \le \gamma, tfin_{last} < tfin_{st} / \theta \\ tfin_{st}, & tfch \le \gamma, tfin_{last} \ge tfin_{st} / \theta \\ tfin_{st} \times pmin, & others \end{cases} \quad (4)$$

Each node only records the latest n times of awareness results. The traffic records are updated according to (5), where $flow_{n-1}$ and $flow_n$ represent two adjacent traffic records, θ is used to adjust the update rate and $0<\theta<1$.

$$flow_n = \theta \times flow_n + (1-\theta) \times flow_{n-1} \quad (5)$$

C. Distributed Topology Decision-making Module

On the basis of the traffic, power consumption and VCG mechanism, the module in each node configures the links connecting the node and devises sleeping and awakening strategies for the links by an online processing way.

Comprehensive traffic value trf^*_{ij} is calculated in (6), where trf_{ij} represents the current link traffic, $trf^{'}_{ij}$ represents the prediction traffic of link, $\beta \in (0,1)$. Let trf^*_{ij} instead of $trf^{'}_{ij}$, then the resulting link value is the link benefit m_{ij}. The link cost indicating the power consumption of active link is calculated in (7), where $power_i$ represents the power consumption of link i, i is a link on the path l, trf^*_i represents a comprehensive traffic value on the link i, c_l represents the cost of path l.

$$trf^*_{ij} = trf_{ij}(1 + \alpha(trf_{ij} - trf^{'}_{ij})(trf_{ij} - trf^{'}_{ij})^{\beta} / \left| trf_{ij} - trf^{'}_{ij} \right|) \quad (6)$$

$$c_l = \sum_{i \in l} power_i / trf^*_i \quad (7)$$

The utility function is defined in (8), where u_i, v_i and p_i represent respectively the utility function, value function and paying function of path i; the path value function is defined in (9), where V_{jk}, m_{jk} and c_{jk} represent respectively the value function, path benefit and path cost between the node j and k; the value function of node v_j is defined in (10), where A_j represents the set of nodes connecting to node j; the path paying function is defined in (11), where c_{jk}^* represents minimal cost path between node j and k, v_j^{-i} represents the value of node j without including node i, m_{jk}^{-i} and c_{jk}^{-i} represent respectively the benefit value and the cost value of the shortest one in the paths including node j and k but without going through i and further the node paying function is defined in (12).

$$u_i = v_i - p_i \tag{8}$$

$$V_{jk} = m_{jk} - c_{jk} \tag{9}$$

$$v_j = \sum_{k \in A_j} m_{jk} + \sum_{k \in N \setminus A_j} (m_{jk} - c_{jk}) \tag{10}$$

$$p_{ji} = \sum_{k \in A_j} m_{jk} + \sum_{k \in R_j \setminus A_j} (m_{jk} - c_{jk}^*) - \sum_{k \in A_j} m_{jk}^{-i} - \sum_{k \in R_j \setminus A_j} (m_{jk}^{-i} - c_{jk}^{-i}) \tag{11}$$

$$P_i = \sum_{j \in A_i} P_{ji} \tag{12}$$

Distributed topology decision-making module consists of two modules as follows.

The VCG operation module operates as follows.

Step1: Taking node i as the source, calculate the path benefit matrix $[m_{ik}]$ and path cost matrix $[c_{ik}]$.

Step2: Send $[c_{ik}]$ and the sub-lowest path cost matrix $[c2_{ik}]$ to the neighbor nodes.

Step3: When neighbor node j receives messages, calculate as follows for any non-neighbor node k of node i.

Step3.1: Calculate the lowest path cost c_1 and the sub-lowest path cost c_2 from node j via i to node k. If c_2 is lower than the original path cost from j to k, then take i as the next hop of j in path jk.

Step3.2: $c_{jk}^* = \min(c_1, c_{jk})$. Select a new c_2 from c_{jk}^*, c_1, c_2 and c_{jk}, and ensure its next hop is different from the former c_2. Update and record.

Step4: For any neighbor node t of the node j, update the paying value: if t is a bottleneck node, then $p_{jt} = p_{jt} + m_{jk} - c_{jk}^*$; if t is a non-bottleneck node, then $p_{jt} = p_{jt} + c_2 - c_{jk}^*$.

The sleeping-awakening module operates as follows.

Step1: Calculate new link paying value p_{ij} according to the paying value update algorithm; $u_{ij} = v_{ij} - p_{ij}$, and if $u_{ij} < 0$, then put the link ij to sleep.

Step2: Calculate the least sleeping time $tmhb$ in (13), where $tmch$ represents the status switching time of awakening a port, $pwhb$ represents the standby power consumption of the sleeping port, $pwch$ represents the instantaneous power consumption of awakening a port, $pwwk$ represents the power consumption of a port operating under a normal case, $0<\theta<1$.

$$tmhb = (tmch \times (pwch - pwhb)) / (\theta \times pwwk - pwhb) \tag{13}$$

Step3: Call the sleeping and awakening module to manage the sleeping link.

Step4: Update the path benefit value m_{ij} and path cost value c_{ij}. c_{ij} is calculated in (14), where t_{ij} represents the sleeping time of link ij.

$$c_{ij} = \begin{cases} \infty, & t_{ij} < tmhb \\ m_{ij}^{-\alpha \cdot (t_{ij} - tmhb + 1)}, & t_{ij} \geq tmhb \end{cases} \tag{14}$$

Step5: If any link related to the node connecting to link ij overloads, enforce to orderly awaken the sleeping link according to the sleeping time from long to short until the overload disappears.

D. Sleeping Control Module

The model is responsible for putting the devices to sleep or awaken, sending, receiving and handling the related sleeping or awaking information, and sending the current link states to the topology decision-making module.

IV. Simulation and Performance Evaluation

A. Simulation

We adopt three kinds of link traffic models shown in Fig. 3. We take the topologies in Fig. 4 of China Education and Research Network (CERNET), CERNET2, NSFNET and INTERNET2 as simulation use cases [11]. The benchmark algorithm in our simulation is the GRiDA algorithm [12].

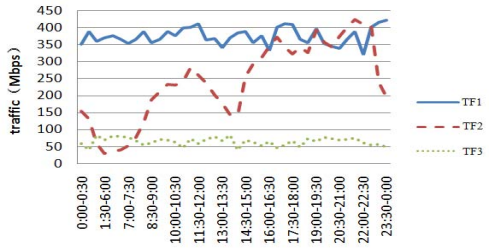
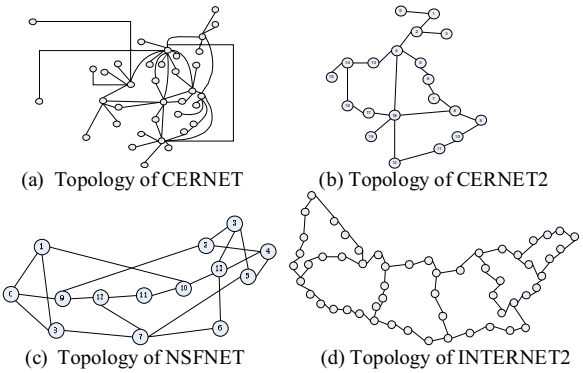


Figure 3. Traffic use cases



(a) Topology of CERNET (b) Topology of CERNET2

(c) Topology of NSFNET (d) Topology of INTERNET2

Figure 4. Topology use cases

B. *Performance Evaluation*

(1) Comparison on power consumption

Periods of regularly collecting information in DPTMS, and GRiDA are 20sec. In Fig. 5, DPTMS shows better power saving and adaptability for various topologies and various traffic models. This is on account of that the traffic prediction module of DPTMS can avoid some decision-making mistakes caused by traffic fluctuation, and the ability of automatically adjusting information awareness interval of DPTMS makes it update data more timely than GRiDA.

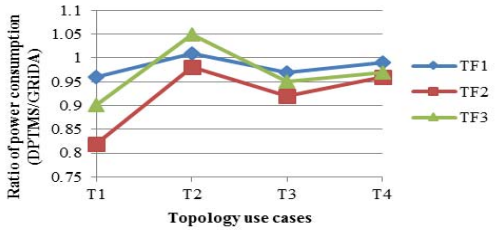


Figure 5. Comparison of power consumption

(2) Comparison on network performance

We take TF2 as the test model and observe the average hops between nodes. The comparison on average path hops is shown in Fig. 6. Average hops in GRiDA are more than that in DPTMS. This is because GRiDA completely depends on the interaction between neighbors without an overall coordination. Further, the differences of the average hops are smaller for the topologies with larger scale and higher complexity, which is mainly due to such topologies lead to more paths to select and a better fault tolerance.

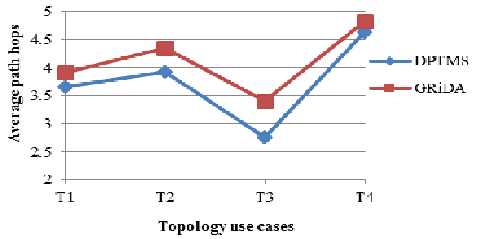


Figure 6. Average path hops

V. CONCLUSION

DPTMS relying on the cooperation of above modules can manage the network topology in real time by an online processing way. Especially, the VCG mechanism is brought into distributed topology decision-making module, which effectively reduces the risks of performance degradation and management scheme collapse. In the simulation, DPTMS shows some relatively satisfying results as compared with GRiDA and indicates its effectiveness and an outstanding power-saving potential.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation for Distinguished Young Scholars of China under Grant No. 61225012 and No. 71325002; the Specialized Research Fund of the Doctoral Program of Higher Education for the Priority Development Areas under Grant No. 20120042130003; Liaoning BaiQianWan Talents Program under Grant No. 2013921068.

REFERENCES

[1] B. G. Bathula, M. H. Jaafar, Elmirghani, "Green networks: energy efficient design for optical networks," Proceedings of the IFIP International Conference on Wireless and Optical Communications Networks (WOCN 2009), IEEE, April 2009, pp. 1-5, doi: 10.1109/WOCN.2009.5010573.

[2] S. Nedevschi, L. Popa, G. Iannaccone, et al, "Reducing network energy consumption via sleeping and rate-adaptation," Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI'08), San Francisco, California, April 2008, pp. 323-336.

[3] A. Cianfrani, V. Eramo, M. Listanti, et al, "An energy saving routing algorithm for a green OSPF protocol," Proceedings of the INFOCOM IEEE Conference on Computer Communications Workshops (INFCOMW 2010), IEEE, March 2010, pp. 1-5, doi: 10.1109/INFCOMW.2010.5466646.

[4] M. Rodriguez-Perez, S. Herreria-Alonso, M. Fernandez-Veiga, et al, "Improved opportunistic sleeping algorithms for LAN switches," Proceedings of Global Telecommunications Conference (GLOBECOM 2009), IEEE, Nov.-Dec. 2009, pp. 1-6, doi: 10.1109/GLOCOM.2009.5425710.

[5] L. Chiaraviglio, M. Mellia, F. Neri, "Reducing power consumption in backbone networks," Proceedings of the IEEE International Conference on Communications (ICC2009), IEEE, June 2009, pp. 1-6, doi: 10.1109/ICC.2009.5199404.

[6] L. Chiaraviglio, M. Mellia, F. Neri, "Minimizing isp network energy cost: Formulation and solutions," IEEE/ACM Transactions on Networking, vol. 20, no.2, April 2012, pp. 463-476, doi: 10.1109/TNET.2011.2161487.

[7] A. A. Kist, A. Aldraho, "Dynamic topologies for sustainable and energy efficient traffic routing," Computer Networks, vol. 55, issue 9, June 2011, pp. 2271-2288, doi: 10.1016/j.comnet.2011.03.008.

[8] S. S. W. Lee, P. K. Tseng, A. Chen, "Link weight assignment and loop-free routing table update for link state routing protocols in energy-aware internet," Future Generation Computer Systems, vol. 28, issue 2, February 2012, pp. 437-445, doi: 10.1016/j.future.2011.05.003.

[9] X. W. Wang, H. Cheng, K. Q. Li, et al, "A cross-layer optimization based integrated routing and grooming algorithm for green multi-granularity transport networks," Journal of Parallel and Distributed Computing, vol. 73, issue 6, June 2013, pp. 807-822, doi: 10.1016/j.jpdc.2013.02.010.

[10] X. W. Wang, W. G. Hou, L. Guo, et al, "A new multi-granularity grooming algorithm based on traffic partition in IP over WDM networks," Computer Networks, vol. 55, issue 3, February 2011, pp. 807-821, doi: 10.1016/j.comnet.2010.11.001.

[11] X. W. Wang, H. Cheng, M. Huang, "Multi-robot navigation based QoS routing in self-organizing networks," Engineering Applications of Artificial Intelligence, vol. 26, issue 1, January 2013, pp. 262-272, doi: 10.1016/j.engappai.2012.01.008.

[12] A. P. Bianzino, L. Chiaraviglio, M. Mellia, et al, "Grida: Green distributed algorithm for energy-efficient ip backbone networks," Computer Networks, vol. 56, issue 14, September 2012, pp. 3219-3232, doi: 10.1016/j.comnet.2012.06.011.

DMAODV: A MAODV-Based Multipath Routing Algorithm

Runping Yang
Computer science department
College of Science and Technology
Ningbo University
Ningbo,China
yangrunping@nbu.edu.cn

Xia Sun
Computer science department
College of Science and Technology
Ningbo University
Ningbo,China
sunxia@nbu.edu.cn

***Abstract*—The MAODV (Multicast Ad hoc On-demand Vector routing protocol)shows a smooth performance in light load ad hoc networks. However, packets transmission ratio will decrease continuously with the increasingly scale of multicast group. In this paper we discussed the impact of attacks on MAODV protocol. Based on these analyses, we proposed a routing protocol DMAODV(Distensible Multicast Ad Hoc On Demand) which is based on MAODV. It adds new feature on flooding message control, put forward a kind of "Self restrain flood" to suppress the MAODV GRPH and RREQ broadcast traffic.We made full use of the broadcast feature of wireless medium, and allow all nodes in the network send multicast packets.we analyzed the influence of nodes' moving on routing. The simulattion with OPNET shows that the speed of packets transmission will increase and delay will decrease in DMAODV, and finally .The results justify that this protocol has better performance, less control overhead and higher scalability.**

Keywords:Ad hoc, MulticastRouting, MAODV, OPNET, simulation

I. INTRODUCTION

Mobile Ad hoc network is a multi-hop temporary autonomous system which is made up of a group of mobile devices with wireless transceiver nodes. With the continuous development of wireless communications technology, Mobile Ad hoc Networks has become more and more widely used in military, emergency rescue and other special circumstances for its flexibility and high-speed in networking[1].

Because of those challenges such as dynamic changes of topology, one-way channel, limited bandwidth of wireless transmission, limitations of mobile terminal capabilities, we should shoulder a heavier task of the research on Mobile Ad hoc network.

The rest of this paper is organized as follows. Section 2 reports a review of previous related works especially for MAODV protocol. In section3, we present the improvement method of proposed system and try to demonstrate the feasibility properties. In section 4 the result of experiments has been described. Finally paper concludes in section 5.

II. RELATED WORK

This section summarizes previous works that are closely related to our proposal.

A. *Multicast Routing Protocol*

The basic idea of multicast routing protocol is sending the multicasting packets to every member of multicast group[2]. Major challenges are the nodes' mobility, the possible loop, non-ideal routing, creating on demand, routing refreshing, packet transmission mode (flooding the whole network or spreading limited with the member nodes).

So far, several multicasting routing protocols has been proposed under the environment of mobile Ad hoc network. They are different from path-finding and routing mechanism[3]. According to the method of routing, they are divided into three groups: on-tree multicast routing, on-grid multicast routing and mixed multicast routing.

B. *Multicast Ad-hoc On-Demand Distance Vector Protocol*

The MAODV (Multicast Ad-hoc On-Demand Distance Vector) routing protocol discovers multicast routes on demand using a broadcast route-discovery mechanism[4]. A mobile node originates a Route Request (RREQ) message when it wishes to join a multicast group, or when it has data to send to a multicast group but it does not have a route to that group. Only a member of the desired multicast group may respond to a join RREQ[5].

If the RREQ is not a join request, any node with a fresh enough route (based on group sequence number) to the multicast group may respond. If an intermediate node receives a join RREQ for a multicast group of which it is not a member, or if it receives a RREQ and it does not have a route to that group, it rebroadcasts the RREQ to its neighbors.

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

III. OPTIMIZATION OF MAODV

In this section, the improvements are made on the scalability of MAODV, and it puts forward a scalable multicast routing protocol based on MAODV(Distensible Multicast Ad Hoc On Demand - short Vector Routing Protocol). It adds some new features that can lead to better

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.62

scalability when the scale becomes larger ,the number of group members increases,or the nodes moves faster.

A. the improvement based on flooding message control

In MAODV protocol,the nodes get the information of their neighbor through neighbor's hello packets.These packets can be made full use in the MAODV protocol routing maintenance. We can improve the methods of flooding to suppress traffic through the information of adjacent nodes ,thus it is named as DMAODV (Distensible Multicast Ad Hoc On-Demand Distance Vector Routing Protocol).

This section puts forward a kind of "Self restrain flood" (Self - Pruning flooding) by Hyojun Lim, the flooding methods to suppress the MAODV GRPH and RREQ broadcast traffic[6].

Self-repression flooding algorithm is described as follows:

1)Node vi contains his neighbor node list N (vi) in forwarding broadcast packets.

2)the node that has received the broadcast packets from node vi will check:if Q(v)=N(vj)-N(vi)-{vi} is null.({vi} is the set only containning node vi);

3)If Q (v) is empty, the node vi's neighbor nodes are given a broadcast packet from vi. vj don't need to forward the broadcasting groups; Otherwise, vj forwarding the broadcasting groups.

DMAODV is proposed to improve MAODV by Self-repression flooding algorithm which is applied in GRPH and RREQ packet.The specific approach is to extend GRPH and RREQ header information to include adjacent nodes.After extension GRPH and RREQ header are shown in table 1,table 2.

TABLE I. *GRPH HEADER*

Type（8bits）	**G(1bit)**	**NeighborConut（8bit）**	**HopCount（8bits）**
DEST （32bits）		DestSeqNb （8bits）	
SRC （32bits）		SrcSeqNb （8bits）	
NeighborHop1（32bits）			
… … …			
NeighborHopN（32bits）			

TABLE II. *RREQ HEADER*

Type（8bits）	**G(1bit)**	**NeighborConut（8bit）**	**HopCount（8bits）**
GRPL（32bits）			
GRP（32bits）		GrpSeqNb（8bits）	
NeighborHop1（32bits）			
… … …			
NeighborHopN（32bits）			

when the nodes broadcast RREQ and GRPH messages,it will fill in the header with his neighbor node number and the IP address of the neighbor nodes.when receiving the RREQ and GRPH ,the node will compare the IP addresses in the packet's header and their neighbors' in the list of IP address.if they are not the same,the broadcast packet will be forwarded,on the contrary it will not be forwarded.

B. the improvement based on routing recovery and routing maintenance

MAODV is trying to send unicast packets through the feedback from the MAC layer to detect link disconnect. It happens at this time that the Link has disconnected,and lead to the missing of packages. In the improved protocol--DMAODV,when the node sends packets, the packets must be assigned a current sequence number to display the parameter of current node, so that the neighbor node will build and refresh the routing.The nodes which have effective routing will send a GRPH message Every setting time T (T>GROUP_HELLO_INTERVAL) after the previous business information or any control after the message is sent .Node receives any business information or control message, the adjacent nodes is to create or update the corresponding effective routing.If a node has not received effective routing of adjacent nodes in any business information or control message in the consecutive NT (N > 1) time, the link is considered break.In accordance with the original MAODV protocol , the node that initialize serial number 1 is considered as a team leader to create a multicast tree .

In the process of routing maintenance, each forwarding data routing has "routing expiration time" whose value is equal to the sum of the current time and ACTIVE_ROUT_TIMEOUT .So we can reduce the survival time of fault link by reducing ACTIVE_ROUT_TIMEOUT, and help updating routing periodically.

As a result, we reduce the time to live of the initial value in DMAODV agreement(TTL_START) and ACTIVE_ROUT_TIMEOUT values.

C. The improvement of transmit mode based on multicast packets

DMAODV is designed to allow all nodes in the network to send multicast data packets,and multicast packets is along the multicast tree when they are being broadcast.With respect to the manner of multicast packets to send improvement is mainly manifested in the following areas:the nodes in the network can send multicast packets.

In MAODV, only multicast group members can send multicast group multicast packets[7].This restriction will cause the loss of data packets sent to a multicast group which reduces the packet of packet delivery ratio .At the same time, because the packets will not reach the multicast group , it will cost more bandwidth of those that can reach the multicast group packets (from the multicast group) and then lead to unnecessary packet delay.

MAODV eliminates this limitation and Support each node's multicast communication in the network[8].This must be taken into account that if the source nodes are not the group member how will the data packets reach the every member of multicast group[9].This process has two steps:

Step 1: we use the AODV routing discovery which reaches the Specific nodes and routing maintenance mechanism to find out the route to the multicast group

memberare . this is just actually mechanism of unicast protocol .

Step 2: After the Members of the tree receive multicast data packets ,they spread the data packet to every member of group in the multicast tree . this is mechanism of MAODV protocol .

In DMAODV,We make full use of the broadcast feature of wireless medium, and allow all nodes in the network send multicast packets. Thereby it can reduce the delay when a lot of packets are loss, and improve the packet delivery ratio further.

IV. SIMULATION

We use OPNET modeler10. 0 [10] as a basic platform for simulation in Windws 2000 and design design a simulation environment according to the Ad hoc network background.

A. *the simulation environment*

The node model includes application layer module,routing layer module,MAC module,physical module.The routing layer module is designed for the requirements in the improved and the original MAODV routing protocol .The MAC layer module uses the CSMA/CA of 802.11a.The physical layer is two-way wireless link of FHSS(Frequency Hopping Spread Spectrum).

50 nodes are randomly distributed in the rectangular area of 1000 x 500 m2 .the nodes use Random Way-point Mobile model,the parameters are set as follows:

TABLE III. *OPNET PARAMETERS*

parameters	values
Nodes count	50
Multicast group count	1
Network coverage area	1000×500m2
The simulation time	900s
Nodes move speed	0～20m/s
Data transmission bandwidth	2Mbps
Node communication distance	250m
MAC protocol	802.11

B. *Implement of Core Code*

Here, we describe the core code of the improvments:

- The key data structure

The structure of multicast routing table is as follow:

```
typedef struct
{
 int dest;              //the IP Address of destination node
 int dest_SeqNb;    //the sequence of destination node
int grp_addr ;       //the IP Address of multicast group
 int grp_leader_addr ;//the leader IP Address of multicast
 int grp_SeqNb ;      //the sequence of multicast group
 int grp_leader_HopCount;  //the hop count to multicast group
 int hop_Count;        //hop count
 int last_Hop_Count; /the hop count to destination
 nodeint next_Hop;     //the next hop
 sFifo* listOfPrecursors;  //the list of precursor
double expirationTime;    //the expirationTime
Route_Status_Type status;  //the status of routing table
Evhandle expiration_evt;  //the routing table's triggering event date
} mrt_table;  //Routing table data structure
```

- the process of nodes joining in multicast group

If some nodes send RREQ control messages to ask for joining multicast group, it will call for "maodv_join" function:

```
static void maodv_join(int grp_addr)
{
        mrt_table  *group_rec;
         group_rec = mrt_table_find(grp_addr);
         if(group_rec==NULL)
        {group_rec=mrt_table_insert(grp_addr);}
        if(!group_rec.is_router)
        {/* if there is no route to the multicast group，
        join RREQ request will be initiated*/
        rreq_route_discovery(grp_addr,
          RREQ_JOIN, NULL, 0);}
        else if(!group_rec.is_member)
        {/* if this node has become forward node of
        multicast data packet,it will become the multicast
        group member. */
        group_rec.is_member = 1;}
}
```

V. ANALYSISOF PERFORMANCE

The purpose of simulation of the mobile Ad hoc routing protocol is to analyze the communication performance of MAODV and DMAODV and discuss the related parameter in network comprehensively and deeply.

In the simulation system that we has built in mobile Ad hoc network, the average node depth is 4;link transmission loss rate is randomly distributed between[0,0.5][8];data packet and NACK packet transmission delay obey exponential distribution between [5，30ms].

The relationship between packet delivery ratio and nodes moving speed are shown in In figure1. The transmission ratio will decrease continuously with the increasingly scale of multicast group and network load under the certain nodes moving speed. However the transmission ratio in DMAODV is much better then in MAODV especially in stability as shown in Fig.1. Because the active maintenance[10] is adoped in DMAODV to reduce the the link failure.In DMAODV all nodes can send multicast packets ,thus the packets loss rate will be reduced and the packets transmission ratio will increase further.

The relationship between average latency and multicast group is shown as fig.2.Under the certain moving speed ,the

average latency will increase with the increasing of multicast group network scale and the network load.But as shown in fig.2,DMAODV is better than MAODV , especially when the speed and the load increasing rapidly.Because self-pruning flooding is adopted in controlling messages to decrease the broadcast to decrease the broadcast packets, and then lead to the decrease of delay. At the same time,the ACTIVE_ROUT_TIMEOUT is reduced to decrease the TTL of fault link.Fig.2 shows that when the multicast group become larger and larger, the node move faster and faster,the average latency of MAODV will increases rapidly.while in DMAODV,the average latency maintains stable .

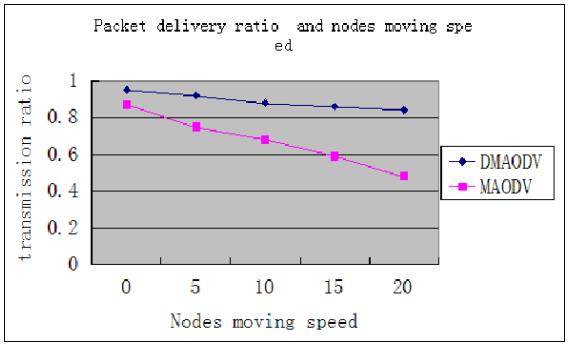
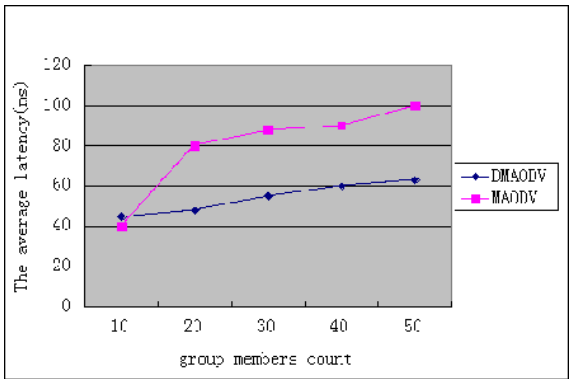


Figure.1 Transmission ratio

Figure.2 The Average latency



VI. CONCLUSION

This article describes the characteristics of mobile Ad hoc networks and focuses on the performance of MAODV multicast routing protocols. On this basis, we proposed improvements to the method of MAODV -DMAODV,and simulate the protocol adopting the OPNET modeler which has acclaimed in the simulation field. Results show that the improved MAODV might provide a viable option for vehicular ad hoc networks with high mobility and density.

Mobile Ad hoc Multicast Routing is the frontier research topics in current network area. As the continuous development of new technologies of internet, there will be some new focus on this the direction of research projects. Our future work will implement and compare other multicasting routing protocol.

ACKNOWLEDGMENT

This work is supported by the Twelfth Five-Year Plan of Zhejiang Province Key Discipline under grand No.20121114,Ningbo Natural Science Fund under Grant No.2013A610003 and the Scientific Research Fund of Zhejiang Provincial Education Department under grant No.Y201016754.

REFERENCES

[1] Hui Xia, Shoujun Xia, Jia Yu,"Applying link stability estimation mechanism to multicast routing in MANETs",Journal of Systems Architecture,Volume 60,ages 467-480

[2] Feng He,Kuan Hao,Hao Ma.S-MAODV,"A trust key computing based secure Multicast Ad-hoc On Demand Vector routing protocol",Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on,2010, pp:434- 438.

[3] Farhat Anwar1, Aisha Hassan Abdalla1, S. M. Sadakatul Bari1."Enhancing Performance of MAODV Routing Protocol for Wireless Mesh Network Using Integrated Multiple Metrics Technique"(IMMT).International Journal of Networks and Communications 2012, 2(2): pp:1-6.

[4] WEI Chengying,DAI Cuiqin,LEI Fang."Multicast Routing Algorithm of Wireless Mesh Network Based on Improved MAODV Protocol",Computer & Digital Engineering,2012,1.

[5] E.Cheng, "On-demand multicast routing in mobile ad hoc networks", M.Eng. thesis, Carleton University, Department of Systems and Computer Engineering, 201,4.

[6] Feng He,Kuan Hao ,Hao Ma.S-MAODV: A trust key computing based secure Multicast Ad-hoc On Demand Vector routing protocol,Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on,2010, pp:434 - 438.

[7] Farhat Anwar1, Aisha Hassan Abdalla1, S. M. Sadakatul Bari1.Enhancing Performance of MAODV Routing Protocol for Wireless Mesh Network Using Integrated Multiple Metrics Technique (IMMT).International Journal of Networks and Communications 2012, 2(2): 1-6.

[8] WEI Chengying,DAI Cuiqin,LEI Fang.Multicast Routing Algorithm of Wireless Mesh Network Based on Improved MAODV Protocol.Computer & Digital Engineering,2012,1

[9] E.Cheng, "On-demand multicast routing in mobile ad hoc networks", M.Eng. thesis, Carleton University, Department of Systems and Computer Engineering, 2012.

[10] Gao Song,OPNET Modeler,Beijing:Electronics Industry Press,2010,9.

Research on Feature Matching of the Field-Based Network Equipment Image

Juan Fang, Qi Yue, Jianhua Wei
College of Computer Science
Beijing University of Technology
Beijing, China
e-mail: fangjuan@bjut.edu.cn,
549568839@qq.com, weijianhua@bjut.edu.cn

Junjie Mao
China Information Technology Security
Evaluation Center
Beijing, China
e-mail: maojunjie@sina.com

***Abstract*—To the problem that the identify of software and hardware information in the reverse analysis of Network equipment firmware image, a firmware header parsing scheme based on matching features of the field has been proposed. Through the analysis of the firmware header got feature values and build signature database, then use matching fields to identify the characteristics firmware header information, and then describes the basic ideas and concrete implementation flow. Experimental results show that the scheme can be obtained analytical results more accurately.**

Keywords-network equipment; firmware image; software and hardware information; feature values; signature database

I. INTRODUCTION

In recent years, with the popularity of the Internet, routers, switches and other network equipment after another, the role of network device for data communications. If these devices are compromised, it will affect all related to the Internet industry, sector. Like an ordinary computer network equipment also includes a processor, operating system, applications, etc., if the network device hidden malicious functionality, or flawed program, most of them exists in the operating system or application of the device. Network equipment firmware image contains the operating system and applications. If you can effectively analyze firmware image can be timely discover the back door, vulnerabilities and other security issues that exist in network devices.

The reverse analysis of network equipment firmware image is that restore the curing executable code format, analyze structure and analyze the implementation process, so you can understand the logic and function of the network device from the software, this is important for detecting vulnerabilities and threats in the equipment [1]. However, the simulation runs of firmware have a large dependence with hardware and software environment of the device [1], so the acquisition of network equipment hardware and software environment is a top priority for the reverse analysis of the firmware image. However, in actual operation, analysts often face unknown binary firmware image, cannot directly determine the operating system version and other relevant information [2]. So the network device firmware image resolution is the premise and foundation of network equipment reverse analysis.

Currently, the research on reverse engineering to network device firmware is less. Zhang Ping in [3] studies the embedded operating system identification technology, proposes a multi-attribute decision based operating system identification scheme. Zhao Ya-xin in [4] proposes an operating system identification scheme that tests the operating system of the device through JTAG interface. Guillaume Delugre and others in [5] reverse analyze the Broadcom Ethernet card firmware, and then prepared a series of basic analytical tools, while adding new features in the original firmware repackaging. These studies aimed at studying a device's properties, and those don't make thorough analysis on all properties of the device of network equipment and cannot meet the needs of reverse engineering.

In this paper, uImage type devices as research object and present a resolution scheme for network equipment firmware image that based on characteristic field matches. By analyzing the meaning of the characteristic field values of the device firmware construct firmware image feature database, then write programs that match judgment feature values of the device firmware image to obtain network equipment hardware and software information, has finished the firmware parsing.

II. RELATED TECHNICAL ANALYSIS

A. Firmware Image Structure

Firmware image is the carrier for operating systems, file systems, applications in embedded device. The content of the firmware was writed by manufacturers when producing [6]. The general formats of the network device firmware image that getted through analyzing the network firmware image is as follows in Table I:

TABLE I. FIRMWARE IMAGE STRUCTURE

Firmware header
BSP
Boot loader
Kernel
Root file system
Applications

Firmware header is the start of a piece of content of the firmware, and indicates the major version number, operating system, architectures and other hardware and software information of the equipment, according to the difference of

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.63

the equipment the header comprises different information. BSP and boot loader is a short code that initializes the hardware and software of the device when the system is starting to make the device work normally. The kernel is the core of the operating system of the device. Root file system is an essential part of the system, including some essential files that make normal operation of equipment. Applications are to achieve some particular function of some equipment [7].

B. Firmware Image Analytical Model

In this paper, the analysis of the network device firmware image is mainly based on structural analysis of the firmware, resolves the information contained by firmware header. First, get different types of feature information of firmware header and add features to the scalable database, then read, process feature values from the firmware header to be parsed and matching judgment feature values with database to parse out the software and hardware information of the device. Firmware image analytical model as shown in Fig. 1:

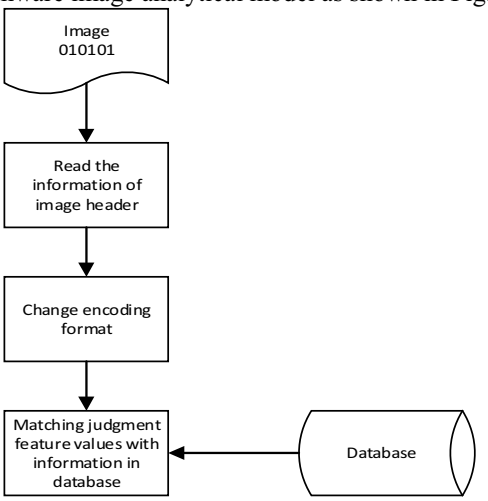


Figure 1. Firmware image analytical model

III. NETWORK EQUIPMENT FIRMWARE IMAGE PARSING

A. The Basic Idea

Features field refers to the content of network device firmware, the content of firmware header can be combined as different data types with certain bytes number, offset data determines the features type represented, such as operating system, kernel architecture, version, device type.

On the basis of the features field, idea the analytical method proposed in this paper is as follows: by operating the target firmware image to be parsed to extract the firmware header content of the firmware image, after transcoding operations for the content of firmware header, get a string of bytes can be used for matching. By interacting with the database, you get the current byte offset and length for the matching operation, and then matched byte string with matching judgment operation, to obtain and output the meaning of the segment byte string.

B. Determine the Feature Value

Through a preliminary analysis to a variety of network device firmware image, we get a series of results, and summarize the results obtained. According to the difference of system's hardware and software platform results were summarized to obtain the corresponding firmware image features values.

This paper focuses on the uImage type network equipment Image, feature values of it were analyzed as shown in Table II:

TABLE II. FEATURE VALUES

Offset Range	Value	Meaning
1-3	0x0519	uImage header, header size: 64 bytes
4-8	X	Header CRC
8-12	X	Created date
12-16	X	Size
16-20	X	Data address
20-24	X	Entry point
24-28	X	Data CRC
28-29	X	OS
29-30	X	CPU
30-31	X	Image type
31-32	X	Compression type

C. Network Devices Firmware Image Feature Database

Due to the current numerous types of network devices and presents growing trend, so an unknown type of network devices firmware image can be appeared to parse. To this end, we designed and implemented a network device firmware image database. Firmware characteristic values of different types of devices were stored in the database and the database provide extended interface, feature values of the emerging device can be easily added to the database used to identify, so the database will be more perfect for the analytical work.

This paper uses SQLite database, SQLite database is a lightweight database, simple and efficient operation, using the database can be well done parsing.

The table of feature values in the database is divided into three levels:

- Table that is composed of first feature value of the firmware header and this is the total table. The first feature value of the firmware header indicates the type of the firmware.
- The eigenvalues table with specific offset and length of a particular type of firmware. The table stores the definitions of all the features field of a particular type of firmware.
- If a feature field has multiple values, and each value represents different meanings, these fields are stored in the table.

This three level eigenvalues table completely stores a particular type of network device firmware image.

IV. NETWORK DEVICE FIRMWARE IMAGE RESOLUTION PROCESS

A. The Overall Process

Network equipment firmware image resolution process is divided into five stages: reading the file content, format transcoding, operating the characteristics database, matching feature information, and outputs the result. If an unknown firmware image file input, process it with these five processes, it can be to output all information that the device firmware included in the header. Specific recognition process as shown in Fig. 2:

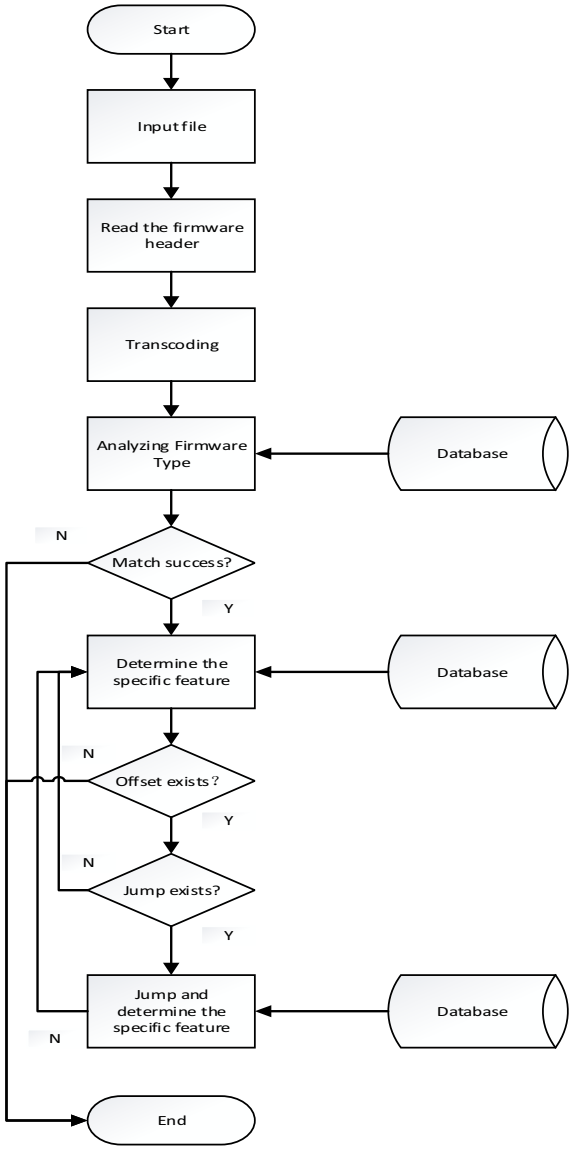


Figure 2. Flow chart

B. Get Features Field To Be Matching

Features field is a string or byte string used to match judgement with the database, is the input used for parsing the information of network device, and only gets the features of the field to be matched before it can be matched with the feature database, so it is the prerequisite and foundation to work normally. The acquisition has the following specific steps:

- Because of network equipment firmware image to be resolved is a binary file, cannot directly use and operation, first read the contents in the program by means of the binary read and save the data block, the block size can be dynamically adjusted as needed.
- The contents of the data block is in accordance with a certain coding method for coding, in order to smooth match the data in the database ,the contents of the data block must be decoded encoding, so that it is same kind of coded data with the data in the database.

C. Feature Field Matches

Feature fields matching in operation is the string to be parsed interact with eigenvalues in the database, according to the three-tier database matching operations are divided into the following steps:

1) Determine the type of network device firmware to be parsed.

a. Open the total table in database.
b. Traversing the next record in the table.
c. Processing each record, to obtain the offset and length of the record.
d. Get eigenvalues from feature string to be parsed by the offset and length.
e. Compared the feature values with feature values of the record, if equal, output its meaning and jumps to the appropriate device type table. Otherwise execute b.

2) Resolve more information on this type of network device.

a. Open the specific device type table.
b. Traverse the next record of the device type table.
c. Get eigenvalues from feature string to be parsed by the offset and length.
d. When the record exist, if the option has jump value, jump to the child table of the record, else output the meaning of the record and execute b.
e. Open the child table of the record.
f. Traverse the next record of the child table.
g. Generation feature value to be resolved according to the offset and length obtained by step c and match characteristic with values of the record obtained by step f, if equal, output the meaning of record obtained by step f and jump to step b. Otherwise, jump to step f.

Through the above process, we can resolve a binary input of unknown network device firmware image file type, it can obtain device hardware and software information that the network equipment binary image file contains.

V. SYSTEM EXPERIMENTS AND RESULTS ANALYSIS

In order to verify the correctness of the program, select the firmware U_W311R_H1_V3.4.0a_CN_20130719092133.bin as the test data for testing.

1) The parameters of the firmware image

As a test object, we have acquired its parameters in Table III.

TABLE III. FIRMWARE PARAMETERS

Firmware	OS	CPU	Image type	Compression
uImage header, header size: 64 bytes	LINUX	MIPS	OS Kernel Image	lzma

2) Characteristics of the field

We try to read and display the contents of the firmware header, because of the windows encoding format, we convert it to GB18030 to display and although it is not the format to match. The feature value is as shown in Fig. 3.

Figure 3. Feature values

3) Running results

We use Python to implement all the features of the system, we call the script to start system and pass the image file to the system, and the result is in Fig. 4.

Figure 4. Running results

According to the results in Figure 4, we know this firmware image has been resolved successfully. It tells us that the type, operating system, kernel architecture, compressed format of firmware, we can use these information to simulated run firmware image.

Due to space limitations, this paper only uses this equipment for testing and gets some information about hardware and software of the network device. The results verify the feasibility of the scheme.

VI. CONCLUSION

The scheme features of the field-based proposed in this paper for network equipment firmware image resolution apply the thought to the network device firmware resolution process, solves the problem that how to configure hardware and software environment when unknown firmware image simulation running. By matching the characteristics field of an unknown binary firmware image, obtaining the meaning of each characteristic field and obtaining the information contained in the firmware header. Experimental results show that the scheme can accurately identify the number of software and hardware information of network equipment firmware image and has higher value. Future research directions are: looking for more features of the field of network equipment, summarized and added to the feature database to constantly improve the analytical capacity of the system to provide more valuable information for the study of network device firmware

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China under Grant No. 61202076.The authors would like to thank the reviewers for their efforts and for providing helpful suggestions that have led to several important improvements in our work. We would also like to thank all the teachers and students in our laboratory for helpful discussions.

REFERENCES

[1] Xie, F., Wu, M. H., Zhang, Z. R. and Ge, Z. H., “Implementation of smart home terminal based on OpenWrt,” Applied Mechanics and Materials, vol. 519, May 2014, pp. 516-519.

[2] GUAN Tao,LU Pei-zhong and LI Dong, “Fine feature for SDH network device identification based on pointer adjustment,” Journal of PLA University of Science and Technology(Nutural Science Edition), vol. 16,No. 1, February 2015,pp. 23–28,doi:10.7666/j.issn.1009-3443.201409006.

[3] ZHANG Ping,JIANG Lie-hui,LIU Tie-ming and XIE Yao-bin, “Embedded systems recognition based on multi-attribute decision making,” Jorunal of Computer Applications, vol. 32,April 2012,pp. 1060–1063, doi:10.3724/SP.J.1087.2012.01060

[4] ZHAO Ya-xin, GUO Yu-dong, and SHU Hui, “Analysis technology of embedded device firmware based on JTAG,” COMPUTER ENGINEERING AND DESIGN, vol. 35,No. 10, October 2014,pp. 3410–3415.

[5] GUILLAUME DELUGRE, “Reverse Engineering the Broadcom NetExtreme's firmware,” Proceedings of HACK.LU '10, Nov. 2010.

[6] Zaddach, Jonas, and Andrei Costin, “Embedded devices security and firmware reverse engineering,” Black-Hat USA, 2013.

[7] Meadows, Cisco Router and Switch Forensics: Investigating and Anal yzing Malicious Network Activity, Elsevier Health Sciences, 2009.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Game Based Multi-domain Protection Scheme in WDM Optical Network

Renzheng Wang, Xingwei Wang, and Min Huang
College of Information Science & Engineering
Northeastern University
Shenyang, China
wangrenzheng@126.com, wangxw@mail.neu.edu.cn, isemhuang@mail.neu.edu.cn

***Abstract*—In order to provide good quality of survivability in the WDM (Wavelength Division Multiplexing) optical network and keep the network load in a good condition, a game based muti-domain protection scheme is proposed in this paper. Game theory, Nash equilibrium and Pareto Optimality of micro-economics are introduced in this shceme for considering benefits of both users and network providers to guarantee the reasonability of survivable routing and achieve all-win. Finally, the proposed protection scheme is implemented by simulation, and the experiment results show that the proposed scheme can reach these design goals.**

Keywords-WDM optical network; multi-domain; game; protection

I. INTRODUCTION

Along with the explosive growth of the Internet traffic, the demand for data transmission bandwidth is higher and higher. For the reason that the WDM (Wavelength Division Multiplexing) technology can provide huge transmission bandwidth, it has become the core technology of the next generation backbone network [1]. Though the WDN technology has improved the transmission capacity greatly, once the network fault occurs, the damage is still huge. It may cause big economic losses. Besides, the scale of network are expanding continuously, the coexistence of multiple network operators results in that the real optical backbone network consists of multiple interconnected domains [2]. For above consideration, we propose a game based multi-domain protection scheme in WDM optical network, in which double-fault cases are considered and the method of segment protection is adopted. For considering benefits of both users and network providers, game theory, Nash equilibrium and Pareto optimality of micro-economics are also introduced to achieve all-win. Finally, the simulation result shows the proposed scheme is feasible and effective.

II. RELATED WORK

Reference [3] proposed a domain-oriented multi-domain protection scheme, in which various protection methods are ranked according to the appropriateness in descending order, when the survivable route needs to be constructed, the protection method ranked in front is selected preferentially. Reference [4] proposed a pre-configured polyhedron based protection method against random multi-link failures in optical mesh networks and it can provide effective protection for random link failures that occur simultaneously or sequentially in network and also improve the utilization of network resources. Reference [5] proposed an end-to-end path protection method under the consideration of four waves mixing in WDM optical networks, in which optical connections that produce error data are regarded as fault links to ensure that the network connection is constituted by the high quality links. Reference [6] proposed a green multicast grooming protection algorithm in the WDM optical network based on two auxiliary graphs to decrease blocking probability and increase bandwidth utilization ratio.

III. PROBLEM DESCRIPTIONS

A. Network Model and Routing Request

Multi-domain network is denoted by $G(V,E,W)$, where V is the node set, E is the physical link set, W is the wavelength set of each physical link. The network is consisted of N domains, $G_i(V_i,E_i,W), i=1,2,\ldots,N$. Routing request is described as $R\left(lp, del_{\max}, tri_{\max}, ber_{\max}, pay_{\max}, lv_p\right)$, where lp is the optical path that needs to be protected, $del_{\max}$ is the maximum delay, $tri_{\max}$ is the maximum transmission impairment, $ber_{\max}$ is the maximum bit error rate, $pay_{\max}$ is the maximum payment, and lv_p is the protection level.

B. Wavelength Routing Graph

In this paper, we improve the layered-graph in [7]. The improved graph is called the wavelength routing graph, where boundary nodes have the wavelength conversion ability and nodes inside the domain do not have such ability. Wavelength nodes in the same layer are connected with the wavelength link. Two layers are connected by the wavelength conversion link. We duplicate nodes in each domain to construct the access layer. The link from access node to wavelength node of the same physical node is called out-access link and the reverse link is called in-access link.

C. Pricing Strategy

The cost-plus pricing method in reference [3] is adopted in this paper. Pricing for wavelength link e^w is as

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.64

$$price_{e^w} = \cos t_{e^w}(1+r) \quad (1)$$

where $\cos t_{e^w} = \cos t_e / |W|$, which is the cost of e^w, and r is the bonus ratio. $\cos t_e$ is the cost of the physical link e.

Situations of wavelength link can be classified as idle, used in working link, used in shared protection and used in dedicated protection. When the wavelength link is shared by different users, the cost of wavelength link is adjusted as

$$\cos t_{e^w} = d \times \cos t_e / |W| \quad (2)$$

where d is the discount ratio when the link is shared. Pricing of wavelength link e^w is as

$$price_{e^w} = \begin{cases} \cos t_{e^w}(1+r) & \text{idle} \\ \cos t_{e^w}(1+r)/n_{share} & \text{used in shared protection} \\ \infty & \text{used in working path} \\ \infty & \text{used in dedicated protection} \end{cases} \quad (3)$$

Prices of the wavelength conversion link, the out-access link of the source node and the in-access link of the destnation are as 0. Pricing for other access links is set as ∞.

D. Game Strategy Set

The strategy set of the network provider is $Play_n = \{dedicated, shared\}$, and the strategy set of the user is $Play_u = \{single, double\}$, which can construct following four protection schemes.

1) Single dedicated protection

Single dedicated protection means one survivable path is constructed and the resource it occupies can not be shared

a) The calculation of survivable path

Step 1: Set prices of links according the pricing strategy.

Step 2: Construct the routing request, which is expressed as $R(v_s', v_d', del_{\max}, tri_{\max}, ber_{\max}, pay_{\max})$, where v_s' and v_d' represent the access node at v_s and v_d respectively, and $del_{\max}$, $tri_{\max}$, $ber_{\max}$ and $pay_{\max}$ are the same as the value in protection request. The route of request R is calculated with the game based routing scheme [8].

Step 3: If routing fails, the single dedicated protection fails; otherwise, discard the access link at both ends of the path, and the rest path is denoted as P_{sd}. Then calculate psr_{sd}, if $psr_{sd} \geq psr_{\min}$, the protection succeeds.

b) The calculation of protection success rate

The single dedicated protection provides a protection success rate of 100% for the single link fault. However, when two links have faults, and one is on the working path and another is on the survivable path, the protection can not be provided. The number of links is denoted as E, the number of hops of the working path is denoted as m, the number of hops that the survivable path passes is denoted as n. The protection success rate of the single dedicated protection is as (6), where C_x^y is the combinatorial number.

$$psr_{sd} = 1 - C_m^1 \times C_n^1 / C_{|E|}^2 \quad (6)$$

2) Double dedicated protection.

Double dedicated protection means two survivable paths are constructed, and resources they occupy can not be shared.

a) The calculation of survivable path

Step 1: Do as Step 1 in single dedicated protection.

Step 2: Do as Step 2 in single dedicated protection.

Step 3: If the routing succeeds, discard the access link at both ends of the route, and the rest route is denoted as P_{dd}^1, which is regarded as the first survivable path. Otherwise, the double dedicated protection fails.

Step 4: Calculate the second survivable path.

Step 5: Set prices of all wavelength links on physical links that P_{dd}^1 passes as ∞.

Step 6: Construct the routing request as the same as the one in Step 2 and calculate the route for the request with the game based routing scheme [8].

Step 7: If the calculation succeeds, discard the access link at the both ends of the path, and the rest path is denoted as P_{dd}^2, which is regarded as the second survivable path. Otherwise, the double dedicated protection fails.

b) The calculation of protection success rate

Because those three links are separated physically, the suceess rate here is 100%, that is to say $psr_{dd} = 1$.

3) Single shared protection

Single shared protection means that one survivable path is constructed, and the resource it occupies can be shared.

a) The calculation of survivable path

Step 1: Set prices of links according the pricing strategy.

Step 2: Do as Step 2 in single dedicated protection

Step 3: If routing fails, the single shared protection fails, otherwise, discard the access link at the both ends of the route, and the rest route is denoted as P_{ss}. Then calculate psr_{ss}, if $psr_{ss} \geq psr_{\min}$, the protection succeeds.

b) The calculation of protection success rate

The single shared protection provides a protection success rate of 100% for the single link fault. When two faults occur on the working path and the survivable path respectively, the protection can not be provided. For all the wavelength links in the survivable path P_{ss}, calculate the union set of physical links of working paths that each wavelength link protects, which is denoted as A. When the working path fails, and another fault link is in A, the protection can not be provided, the success rate is as

$$psr_{ss} = 1 - C_m^1 \times C_{n+|A|}^1 / C_{|E|}^2 \quad . \quad (7)$$

4) Double shared protection

Double shared protection means two survivable paths are constructed, and resources they occupy can be shared.

a) The calculation of survivable path

Step 1: Set prices of links according the pricing strategy.

Step 2: Do as Step 2 in single dedicated protection.

Step 3: If the calculation succeeds, discard the access link at the both ends of the path, and the rest path is denoted as P_{ds}^1, which is regarded as the first survivable path. Otherwise, the double shared protection fails.

Step 4: Calculate the second survivable path.

Step 5: Set prices of all wavelength links on physical links that P_{ds}^1 passes as ∞.

Step 6: Do as Step 6 in double dedicated protection.

Step 7: If the calculation succeeds, discard the access link at the both ends of the path, and the rest path is denoted as P_{ds}^2, which is regarded as the second survivable path. Otherwise, the double shared protection fails.

b) The calculation of protection success rate

Like the double dedicated protection, $psr_{ds} = 1$.

E. Utility Computing

When conducting the game, utilities of the user and the network provider need to be calculated for each strategy combination (f_i, g_j), where $f_i \in Ploy_n$, $g_j \in Ploy_u$.

1) User Utility

User utility is defined as

$$uu_{ij} = \alpha_{pay} \times \left(pay_{\max} - pay_{ij} \right) \Big/ pay_{\max} + \alpha_{pst} \times F \Big/ pst_{ij} \quad (8)$$

where pay_{ij} is the actual payment decided through the game based routing scheme, pst_{ij} is the protection switch time, and α_{pay} and α_{pst} are weights. There are several time parameters, such as fault detection time F, node delay D, link transmission delay P, node configuration delay, C. Then, the protection switch time is calculated as follows.

For single dedicated protection, when the network detects a fault in working path, the protection switch time is as

$$pst_{sd} = F + (m-1) \cdot P + m \cdot D + 2 \cdot n \cdot P + 2 \cdot (n+1) \cdot D \quad (9)$$

For double dedicated protection, two situations exist.

Situation 1: two faults both occur in the working path.

$$pst_{dd} = F + (m-2) \cdot P + (m-1) \cdot D + 2 \cdot n_1 \cdot P + 2 \cdot (n_1+1) \cdot D \quad (10)$$

where n_i is the number of hops of the i th survivable path.

Situation 2: two faults occur on the working path and the first survivable path respectively.

$$\begin{aligned} pst_{dd} = {} & F + (m-1) \cdot P + m \cdot D + 2 \cdot (n_1 - 1) \cdot P \\ & + 2 \cdot n_1 \cdot D + 2 \cdot n_2 \cdot P + 2 \cdot (n_2 + 1) \cdot D \end{aligned} \quad (11)$$

The protection switch process of the single shared protection is similar to that of the single dedicated protection. But in the shared protection, one wavelength link may be shared by multiple survivable paths, thus the node should be configured first, which will bring the delay. So, the protection switch time of the single shared protection is as

$$pst_{ss} = F(m-1)P + m \cdot D + 2n \cdot P + 2(n+1)D + (n+1)C \quad (12)$$

The protection switch process of the double shared protection is similar to that of the double dedicated protection, except the node configuration delay.

Situation 1: two faults both occur in the working path.

$$\begin{aligned} pst_{ds} = {} & F + (m-2) \cdot P + (m-1) \cdot D \\ & + 2 \cdot n_1 \cdot P + 2 \cdot (n_1 + 1) \cdot D + (n_1 + 1) \cdot C \end{aligned} \quad (13)$$

Situation 2: two faults occur on the working path and the first survivable path respectively.

$$\begin{aligned} pst_{dd} = {} & F + (m-1) \cdot P + m \cdot D + 2 \cdot (n_1 - 1) \cdot P + 2 \cdot n_1 \cdot D \\ & + 2 \cdot n_2 \cdot P + 2 \cdot (n_2 + 1) \cdot D + 2 \cdot n_1 \cdot C + (n_2 + 1) \cdot C \end{aligned} \quad (144)$$

2) Network Provider Utility

The network provider utility is defined as

$$nu_i = \beta_{pay} \cdot \left(pay_{ij} - \cos t_{ij} \right) \Big/ pay_{ij} + \beta_{new} \cdot w_{ij}^{new} \Big/ n_{ij}^{p} \quad (15)$$

where $\cos t_{ij}$ is the cost, w_{ij}^{new} and n_{ij}^{p} is the number of newly assigned wavelength links and total wavelength links of the survivable path, and β_{pay} and β_{new} are weights

F. Game Analysis

In the game based protection scheme, four protection schemes construct a utility matrix $R_{2\times 2} = [uu_{ij}, nu_{ij}]_{2\times 2}$. uu_{ij} and nu_{ij} are utilities of users and network providers under the strategy combination $\left(f_i, g_j \right)$. The objective of the game is to find the strategy combination, which can make utilities of both sides optimal. If both uu_{ij} and nu_{ij} is bigger than others, the corresponding strategy combination reaches Nash Equilibrium. But when multiple such solutions exist, Pareto dominant (16) of them need to be compared.

$$pareto_{ij} = 1 \Big/ \left(\eta_1 / uu_{ij} + \eta_2 / nu_{ij} \right) \quad (16)$$

where η_1 and η_2 are weights of corresponding items.

IV. ALGORITHM DESIGN

Step 1: According to the protection request, calculate survivable paths corresponding to four protection strategies.

Step 2: For each successful protection strategy, calculate utilities of the user and network provider with (8) and (15).

Step 3: If four protection strategies all fail, the game based protection fails and the algorithm ends.

Step 4: Check the number of Nash equilibrium solutions.

Step 5: If there is only one Nash equilibrium solution, this solution is regarded as the protection solution.

Step 6: If zero or multiple Nash equilibrium solutions exist, calculate Pareto dominants and choose the best one.

V. Simulation and Evaluation

NSFNET, GEANT and REDIRIS are chosen as three domains. Each domain has 6 boundary nodes, and 3 inter-domain links exist between each two domains. Enhanced Local Segment-Shared Protection (ELSSP) [9] is chosen as the benchmark. The number of business requests obeys Poisson distribution with the mean β. The time interval between two business requests obeys negative exponential distribution with the mean $1/\beta$. Duration of the business obeys negative exponential distribution with the mean $1/\mu$. Thus β/μ (Erlang) is the traffic intensity. The number of wavelengths in each physical link is 16. The number of logical sub links in each logical link is 2. The granularity of business request is one wavelength. For other parameters, we set $\alpha_{pay}=0.5$, $\alpha_{pst}=0.5$, $\beta_{pay}=0.5$, $\beta_{new}=0.5$, $\eta_1=0.5$, $\eta_2=0.5$, $F=10\mu s$, $D=10\mu s$, $P=400\mu s$ and $C=10\mu s$.

As shown Fig. 1, when the traffic intensity is low, the ratio of choosing the dedicated protection strategy is high, thus GSP (Game based Segment Protection) has a smaller MPST (Mean Protection Switch Time). When the traffic intensity is high, GSP can provide double protection, thus GSP has a larger MPST. As shown in Fig. 2, GSP has a higher user utility, because GSP can provide the double protection, but ELSSP only provides the single protection. As shown in Fig. 3, GSP has a higher network provider utility, because GSP can choose the protection strategy according to the network resource utilization and the profit.

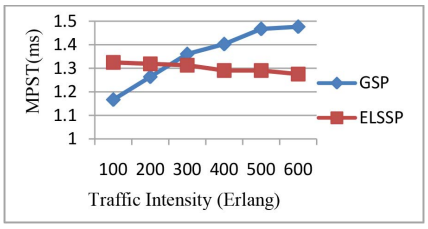


Figure 1. Curve of MPST

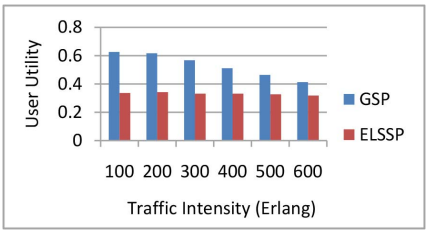


Figure 2. Comparison of the User Utility

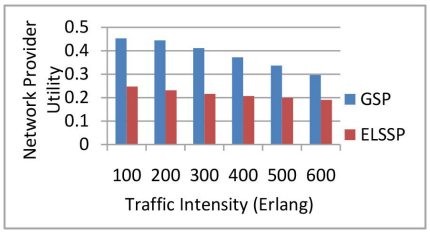


Figure 3. Comparison of the Network Provider Utility

VI. Conclusions

This paper proposes a game based multi-domain protection scheme in WDM optical network, considering utilities of both users and network providers. The scheme can ensure that users can get satisfactory survivability quality with acceptable payment, and network providers can get satisfactory profits from the survivability service and keep the network load in a good condition. In the end, the evaluation result shows that such game based multi-domain protection scheme is effective.

Acknowledgment

This work is supported by the National Science Foundation for Distinguished Young Scholars of China under Grant No. 61225012 and No. 71325002; the Specialized Research Fund of the Doctoral Program of Higher Education for the Priority Development Areas under Grant No. 20120042130003;.Liaoning BaiQianWan Talents Program under Grant No. 2013921068.

References

[1] Mukherjee, B., “WDM optical communication networks: progress and challenges,” Selected Areas in Communications, IEEE Journal on , vol.18, no.10, pp.1810–1824, Oct. 2000

[2] Chalasani, S.; Rajaravivarma, V., “Survivability in optical networks,” System Theory, 2003. Proceedings of the 35th Southeastern Symposium on , pp.6–10, Mar. 2003

[3] Wang, Yue, et al., “A novel multi-domain protection scheme in hybrid optical networks,” Proceedings of SPIE, Nov. 2007.

[4] Shanguo Huang, et al., “Pre-configured polyhedron based protection against multi-link failures in optical mesh networks,” Opt Express, vol. 22, no. 3, pp. 2386–2402, Feb. 2014

[5] Thangaraj J., et al., “End-to-end path protection considering four wave mixing in multi-domain WDM optical networks”, Journal of Optics, vol. 42, no. 3, pp.268–280, Sep. 2013

[6] Yu C., et al., “A new green multicast grooming protection algorithm in WDM optical networks,” Optik-International Journal for Light and Electron Optics, vol. 125, no. 2, pp. 657–662, Jan. 2014

[7] Chen, C.; Banerjee, S., "A new model for optimal routing and wavelength assignment in wavelength division multiplexed optical networks," INFOCOM, Proceedings IEEE, pp.164–171, Mar. 1996

[8] Xinxin Sun, Xingwei Wang, Min Huang, “A Game Based Multi-domain Routing Scheme in WDM Optical Network”, Advanced Computer Architecture, 2014.

[9] Zhang X. N., et al., “On segment-shared protection for dynamic connections in multi-domain optical mesh networks,” International Journal of Electronics and Communications, vol .64, no. 4, pp. 366–371, Nov. 2008

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A real-time micro-sensor upper limb rehabilitation system for post-stroke patients

Guanhong Tao[1,2], Yingfei Sun[1], Zhipei Huang[1], Jiankang Wu[1]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]The 29th Research Institute of China Electronics Technology Group Corporation, Chengdu, China
taoguanhong09@mails.ucas.ac.cn,{yfsun, zhphuang, jkwu}@ucas.ac.cn

***Abstract*—There has been an urgent need for a ubiquitous and quantitative upper limb rehabilitation method for post-stroke patients. This paper presents a real-time micro-sensor-based upper limb rehabilitation system for quantitatively evaluating the patient function status and the rehabilitation training progress. The rehabilitation system mainly consists of three subsystems: a sensor subsystem, a data fusion subsystem, and a rehabilitation training subsystem. The sensor subsystem collects upper limb motion signals and transfers them to the data fusion subsystem. The data fusion subsystem fuses motion signals to obtain motion parameters. The rehabilitation training subsystem visualizes the whole rehabilitation training process in 3D virtual space, visually guiding the patient in the training, highlighting the progress and existing issues. The system can provide rehabilitation ubiquitously, reduce the cost, and bring convenience to patients and families.**

Keywords: Micro-sensor; real-time; rehabilitation system; post-stroke;

I. INTRODUCTION

Cerebral apoplexy, generally known as stroke, is the rapid loss of brain function due to brain disturbance in the blood supply system [1]. According to World Health Organization (WHO), 15 million people suffer stroke worldwide each year. Within China, there are more than 2 million stroke cases each year and there are accumulated more than 7 million patients. Due to limited resources (hospitals, specialists and equipments) the recovery rate is only 30%. Existing post-stroke rehabilitation mainly relies on specialists' manual examination and personal judgments, and the training is under the specialist's supervision. There is always a lack of qualified rehabilitation specialists. Rehabilitation training is a long process, patients and families prefer to be at home or community. Besides, it is painful so that many patients do not have strength to fully cooperate. As such, there is a dying need for a quantitative post-stroke rehabilitation system, which is ubiquitous, intelligent, motivated and immersive.

There are products, such as Vicon [2], using structured high precision video cameras for motion analysis. It is very expensive, requiring special lab space, and not many hospitals can afford. The rapid advances of micro-electro-mechanical systems (MEMS) technology during last two decades have permitted the creation of micro-sensors that are suitable to be placed on human body. In contrast to traditional optical capture systems, micro-sensor motion capture has the advantages of being ubiquitous, simple and low-cost. Freedom from infrastructure requirements allows micro-sensor motion capture to be performed in almost any location.

This paper presents an ambulatory low-cost real-time rehabilitation system, which employs micro-sensors (accelerometers, magnetometers and gyroscopes) to obtain the upper limb motion data of patients and collect a group of motion quality evaluation indicators of universal significance according to the clinical needs in evaluating patient's upper limb motion quality.

The rest of the paper is organized as follows. Section II-IV describe the design of the system, including the hardware design, data processing and rehabilitation methods. And the experimental results will be given in Section V.

II. SENSOR SUBSYSTEM

A. Sensor Type

In the rehabilitation system, each sensor unit consists of a STMicroelectronics L3G4200D [3] tri-axis gyroscope, a LSM303D [4] tri-axis accelerometer and a L3G4200D [3] tri-axis magnetometer. By attaching IMUs onto the surface of human body segments, these sensors can measure acceleration vector, magnetic field vector and rotational rate vector of the body segment. Accordingly, the motion parameters of body segments such as orientation and position can be estimated by fusing sensor measurements. These parameters will be used in various ways to monitor and analyze many aspects of the movement in Section V.

B. Sensor Setup

The sensor subsystem mainly includes two parts: a base station and sensor units. The sensor units are attached to upper limb segments embedded in the fabric of garments. Four micro-sensor units are attached to trunk, upper arm, lower arm and hand, respectively, as shown in Figure 1. Sensor units are wired connected to the base station using shielded cables. The sensor units sample and collect human motion signals, and send them to the base station via I^2C protocol. The base station gathers the motion signals from all the sensor units, packets the data and sends the packages to the data fusion subsystem for further processing. The sampling rate can be adjusted according to applications, and up to 200Hz. Rechargeable Li-ion battery pack is used to provide power for the sensor subsystem. Taking the range of human activity and exercise level of comfort into account, the system uses wireless communications to send data between the base station and the data fusion subsystem. Depending on different circumstances, the system uses two different wireless communication protocols: Bluetooth and

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.65

Wi-Fi, while the former one for indoor environments and the latter one for out-door applications.

Figure 1. Placement of micro-sensors in human upper limb.

III. DATA FUSION SUBSYSTEM

As mentioned in Section II, the data fusion subsystem receives sensor signals from the base station at each sampling time. The signal from each sensor contains three 3-dimensional vectors: acceleration vector, angular velocity vector and magnetic field strength vector. These signals are fused to obtain the orientation and position of each body segment in the data fusion subsystem.

The data fusion subsystem first performs cross-modality fusion, in which the body segment orientation is estimated from the three different modalities of sensors. A quaternion-based Unscented Kalman Filter (UKF) is designed to combine these complementary data sources [5]. Under this filtering framework, the integrated quaternion from gyroscope signals is corrected by the gravity and earth magnetic direction. To optimize the performance under linear acceleration interference, the UKF fuses gyroscope, accelerometer and magnetometer signals by discriminating of gravity and linear acceleration. After subtracting linear acceleration from accelerometer measurement, the corruption of gravity vector observation is minimized. After orientation estimation, the fusion subsystem performs cross-sensor fusion, which estimates position of the segments based on human biomechanical model and forward kinematics. Then, each joint's relative positions are obtained from hierarchical transformations, where the trunk is taken as root, and then is spread to other upper limbs [5].

IV. REHABILITATION TRAINING SUBSYSTEM

In the rehabilitation training subsystem, we develop two types of quantitative evaluation methods: one is based on the existing professional assessment measures, Fugl Meyer assessment [6] of physical performance. By using this measure, suitable rehabilitation schemes are selected from the rehabilitation scheme base. The other is to evaluate the progress of the training, by using the distance measure between the patient's movement/function and the one performed by normal person. This will provide a continuous numerical measure which will be visualized in the 3D immersive and interactive training program to guide the patient and to decide patient's score in the game scenario. These two evaluation methods will be briefly addressed as follows.

A. Fugl Meyer assessment

According to Fugl Meyer assessment [6], five motion quality evaluation indicators of universal significance are proposed for the clinical needs. Explanations of the five evaluation indicators are given as follows.

1) Joint Angle

Joint angle refers to motion arc, which is produced by the movement of joint due to the free contraction of joint muscle. By measuring joint angle, physicians check patients' motion function condition so as to examine the impairment degree of angle motion range. We suppose that upper arm, forearm, and hand are all rigid bodies, rotating around their corresponding joints. Therefore, the joint angle of the shoulder joint is the same as the corresponding angle of upper limb vectors between the start point and the endpoint:

$$\theta = \cos^{-1}\left\langle \begin{matrix} q^{s}_{shoulder} \otimes L^{B}_{upper} \otimes (q^{s}_{shoulder})^{-1} \\ q^{e}_{shoulder} \otimes L^{B}_{upper} \otimes (q^{e}_{shoulder})^{-1} \end{matrix} \right\rangle \quad (1)$$

where $q^{s}_{shoulder}$ and $q^{e}_{shoulder}$ respectively represent the orientation of shoulder joints at the start point and the endpoint. L^{B}_{upper} is the vector of upper limb in body coordination system.

The joint angle of elbow and wrist could also be derived in the similar way.

2) Angle Divergence

Stroke patients usually suffer from upper limb motion disorders and cannot precisely locate the target position. The Angle Divergence (AD) is used to quantify patients' control of trajectory. AD is the angle between the actual measured position of patients` upper limb and the experimental prescribed position of common people. The larger the angle is, the worse the patient's control of motion trajectory will be. AD is calculated as follows:

$$AD = \frac{\sum_{i=1}^{N} \left\{ \begin{matrix} 0, & \text{if } z_i \le z_{i-1} \\ \cos^{-1}\left(\frac{\vec{P}\cdot\vec{T}}{\vec{P}}\right), & \text{if } z_i > z_{i-1} \end{matrix} \right\}}{\sum_{i=1}^{N} \left\{ \begin{matrix} 0, & \text{if } z_i \le z_{i-1} \\ 1, & \text{if } z_i > z_{i-1} \end{matrix} \right\}} \quad (2)$$

where $\vec{P}$ is the vector between the actual measured position and the initial position of joints during the process of motion and $\vec{T}$ is an unidirectional vector in theatrical trajectory. z_i is the joint position of the Z-axis of the i-th sampling time. Due to motion function impairment, the patients often halt and tremble in the process of getting certain objects. Therefore, in calculating AD, only the effective motion will be taken into consideration, that is, only when patients' upper limb is moving to the direction of the object ($z_i < z_{i-1}$), the angle between the vectors of two adjacent points will be calculated.

3) Average Speed

Clinically the efficiency of a patient completing a certain action should be observed. Average speed is adopted here to reflect the efficiency of the patients' completing actions of getting certain objects. The greater the average speed is, the higher the efficiency of completing actions will be. The average speed is the mean value of the instantaneous speed of the respective sampling time, which is shown in the following equation:

$$\bar{v} = \frac{1}{N}\sum_{i}^{N} v(i) \quad (3)$$

where $v(i)$ is the instantaneous speed at the time i, and according to the three-dimensional coordinate values of the two samples before and after the moment, we can calculate the instantaneous speed:

$$v(i) = \frac{1}{2\Delta t}\sqrt{\left(P_x^{i+1} - P_x^{i-1}\right)^2 + \left(P_y^{i+1} - P_y^{i-1}\right)^2 + \left(P_z^{i+1} - P_z^{i-1}\right)^2} \quad (4)$$

where Δt is the sampling interval, P_x^i, P_y^i,P_z^i are the three-dimensional coordinates at the sampling moment of i and the average speed can be derived as:

$$\bar{v} = \frac{1}{2\Delta tN}\sum_{i}^{N}\sqrt{(P_x^{i+1} - P_x^{i-1})^2 + \left(P_y^{i+1} - P_y^{i-1}\right)^2 + (P_z^{i+1} - P_z^{i-1})^2} \quad (5)$$

4) Torso Balance Degree

In the clinical practice, the patients sometimes cannot get objects by controlling their upper limbs and they will instinctively lean forward by using their torsos to finish the task. In the clinical practice, this kind of leaning is judged by observing whether they shrug their shoulders. But this acts as a qualitative method, cannot reflect the extent of the torso forward specifically. Torso balance degree is the angle between the patient's actual torso position and the prescriptive normal people's torso position. Therefore, the patient's torso compensatory condition can be reflected through the torso balance. The greater the value is, the more serious the patient's torso compensatory condition will be. Torso balance is calculated as follows:

$$C_{Torso} = \frac{\max\{\vartheta,\beta,\gamma\}}{90^{\circ}} \cdot 100\% \quad (6)$$

where ϑ, β, γ are the euler angles of torso orientation.

5) Entropy

Tremor is a phenomenon caused by patients' muscle weakness, muscle contraction and relaxation. It is shown as unconscious halt during the process of getting certain objects. Entropy is used here to reflect motion discontinuity caused by tremble, the greater the value which deviates from the normal range is, the more disordered the patient's motion and the worse the tremble will be. The speed is divided into five levels, and the speed gradient of each level is:

$$\tilde{v} = \frac{v_{max}}{step} \quad (7)$$

where v_{max} is the peak speed. Step is the number of divided speed levels. The probability of speed falling into the interval i:

$$P_i = \frac{Number\{\tilde{v}_i < v \le \tilde{v}_{i+1}\}}{N} \quad (8)$$

The numerator is the sampling points which fall into the interval i and the denominator N is the total number of sample points. So entropy is:

$$H = -\sum_{i=1}^{n} P_i \cdot \log(P_i) \quad (9)$$

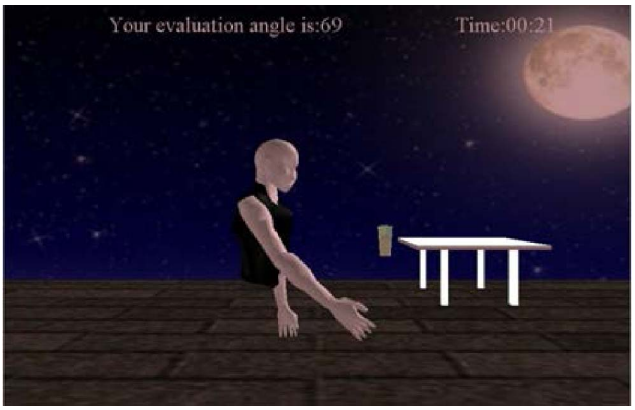


Figure 2. Interactive game of the rehabilitation system.

B. Immersive and interactive training

The above evaluation indicators are used in the rehabilitation game interaction and rehabilitation process management, which digitalizes and visualizes the complex upper limb rehabilitation problem, making the patient's rehabilitation exercise more standardized, targeted and entertaining.

The evaluation indicators are used in game training and developing exercise programs.

1) Evaluation indicators for scenario-based game exercise. Firstly, there will be a motion evaluation of the patients and the evaluation results will be fed back into game exercise (as shown in Figure 2). For example, in the interactive game of getting cups, the position of the cup will be set by the system on the basis of the maximum joint angle which the patient can reach in evaluation. When the patient gets the cup, the system will automatically lift up the cup by 5°. If the patient tries twice but still cannot get the cup, the system will move down by 5°. Thus, the game rehabilitation exercise becomes individualized and targeted. When the patient gets the cup, the system will calculate the patient's angle divergence, average speed, torso balance and entropy. Compared with the evaluation indicators last time, suggestions will be given such as: "Your torso incline angle is too large. Please lean against your chair. You can make it!"
2) The evaluation indicators could be further used to develop exercise programs. Physicians check the patients' evaluation indicators so as to develop targeted exercise programs. For instance, in a patient's evaluation indicators, the evaluation indicator of elbow is relatively poor, so the physician will pay more attention on the patient's motion of elbow joint exercise.

V. EXPERIMENT RESULTS

In the experiment, we respectively measure the getting objects function of normal people and patients with the proposed rehabilitation systems. The subjects of the experiment include 22 adults, who are divided into two groups: patients group with 16 people and healthy group with 6 people, with ages ranging from 40 to 70 years old. The experiment has been approved by the Ethics Committee of Peking University First Hospital and all the subjects have signed informed consent form.

In the experiment, the subjects are asked to get objects orientated in the following two ways:

1) Getting objects with shoulder flexion orientated. The subject sits with his or her back lean against the back of the chair. A cup is settled at the position where the subject's shoulder flexion is 90° and hands could reach, with elbows fully stretched, palm thumbs up.
2) Getting objects with shoulder abduction orientated. The subject sits with his or her back lean against the back of the chair. A cup is settled at the position where the subject's shoulder abduction is 90° and hands could reach, with elbows fully stretched, palms down.

We analyze the evaluation indicators statistically. The evaluation indicators are analyzed in normalized mean value, standard deviation between the healthy group and the patient group of shoulder flexion and shoulder abduction respectively. As can be seen from Figure 3 and Figure 4, joint angles of the patients are generally smaller than that of the healthy people, implying a lack of exercise capacity in patients. Angle divergence of the patients is greater than that of the healthy people, indicating a weaker control of trajectory of patients. Average speed of patients is smaller than that of the healthy people, indicating that patients' lack of exercise capacity has resulted in a low exercise efficiency of the patients. Torso balance of the patients is greater than that of the healthy people, indicating a serious torso compensatory condition of the patient. Entropy of the patient is larger than that of the healthy people, which means that the patients have pretty much cease and tremble.

Given the comparison of Figure 3 and Figure 4, different actions and evaluation indicators have a great distinction between both the patients and the healthy, which means that the five evaluation indicators proposed in the paper have universal applicability in evaluation of motion quality.

VI. CONCLUSION

This paper presents a real-time micro-sensor-based upper limb rehabilitation system for quantitatively evaluating the patient function status and the rehabilitation training progress. Based on the system, physicians have a quantitative evaluation of the patient's motion impairment degree and make up efficient scientific treatment programs.

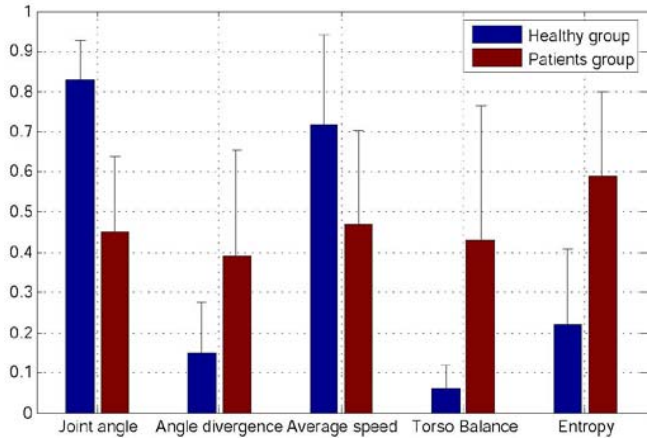


Figure 3. The evaluation indicators mean variance of shoulder flexion motion.

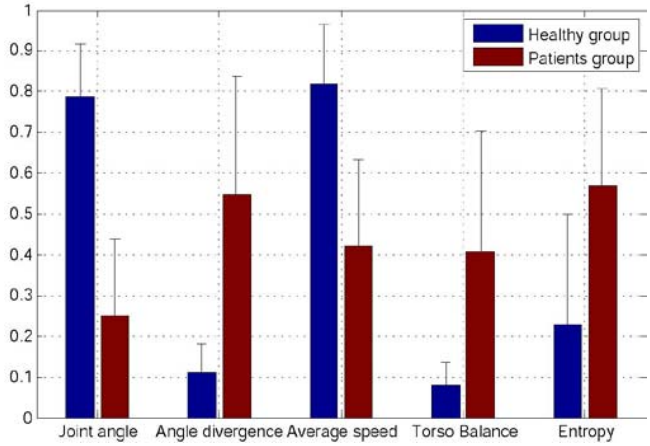


Figure 4. The evaluation indicators mean variance of shoulder abduction motion.

ACKNOWLEDGMENT

This work is supported by the International Science and Technology Cooperation Program of China (2012DFG11820) and National Natural Science Foundation of China (61431017).

REFERENCES

[1] ML Dombovy, BA Sandok and JR BasfordV, "Rahabilitation for Stroke: A Review, Stroke", 1986, pp.363-369.

[2] Vicon Motion Systems Ltd, Vicon life science, http://www.vicon.com/motion-capture/life-sciences.

[3] STMicroelectronics Inc., L3G4200D, MEMS motion sensor: three-axis digital output gyroscope, http://www.st.com/

[4] STMicroelectronics Inc., LSM303D, Ultra compact high performance e-compass: 3D accelerometer and 3D magnetometer, http://www.st.com/

[5] GH Tao, Huang ZP, YF Sun., SY Yao and JK Wu, "Biomechanical model-based multi-sensor motion estimation". Proc. IEEE Sensors Applications Symp. (SAS2013), IEEE Press, Feb. 2013, pp.156-161.

[6] AR. Fugl-Meyer, L. Jaasko, L. Leyman, S. Olsson, and S. Steglind. "The post-stroke hemiplegic patient 1. a method for evaluation of physical performance". Scandinavian Journal of Rehabilitation Medicine, 7(1):13–31,1975.

The network model based on IOCP memory control key technical analysis

Shu Heng
Guizhou Normal University
GuiYang ,China
Shuheng_mast@aliyun.com

***Abstract*—In order to realize the high-speed model of window, we analyze the principle and the mechanism of IOCP. And several key questions of memory management in the solution is given. Through the pressure test, the memory control mechanism can effectively improve the performance of the system, and the memory consumption and the stability of the system are all outstanding.**

Keywords- IOCP ; socket ; memory pool; key technical

I. INTRODUCTION

With the continuous development of computer network technology, in many operators to provide users with efficient and convenient resource retrieval, resource download service process, due to network traffic, security, and intellectual property concerns, need a kind of high efficient and safe way of network communication.[1][3]

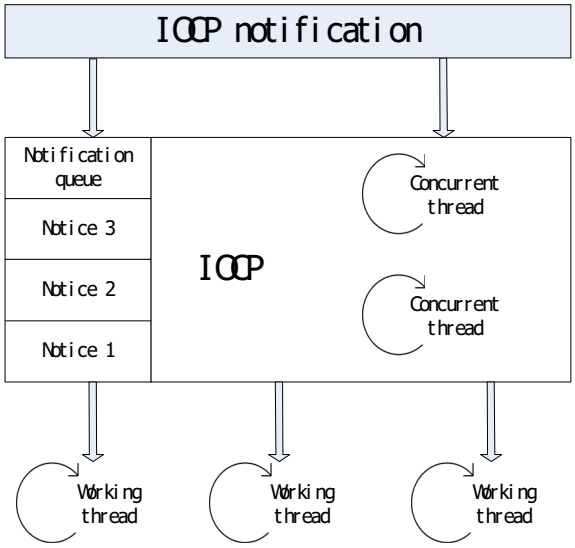
The traditional TCP concurrent server is for each client to create a child threads handle independently. In response to the operation of multiple clients and servers to create such as dry child thread to deal with. And create a thread is a kind of very consume resources, and when multiple clients at the same time to the server for communication, system kernel need to constantly switch threads to run.[2] So the kernel needs to spend a lot of time to transform the running thread context switches, and the allocation of resources. So the communication (synchronous non-blocking) inefficient and also a lot of system resources consuming.

IOCP (I/O Completion Port) is a kind of asynchronous blocking type communication pattern. It can reasonable use and management of multithreading mechanisms, can help to deal with a large number of client requests network service problems, and can make the performance of the system to achieve a better condition. IOCP kernel mode is actually a system issued a notice to the user mode of communication mode. It is by far the most efficient Windows platform of network communication model, and it has a good scalability. Using this model management a large number of sockets, can achieve the best performance of the system.[1][2][3][5]

II. IOCP PRINCIPLE

A. IOCP Principle[2][3][5]

ICOP is a queuing system maintained by the kernel, the operating system of the overlapping IO operation is complete event notification placed in the queue. Then notify thread to handle the appropriate message. After a IOCP creation, it can be associated with multiple file handles. Then overlap IO operations on those file handles. When an IO operation is complete, an overlapping IO completion notification event will be put on this port completion queue, at the same time, a worker thread will be awakened to perform specific processing [1]. The working principle of the diagram below



B. The working process of the IOCP

1.IOCP complete notice [2].

Overlapped I/O operation is completed, the system the complete notice sent to the port. The completion port to maintain a complete notification queue. The operating system has been completed overlapped I/O request notification in the party.

2. The worker thread.[2][4][5]

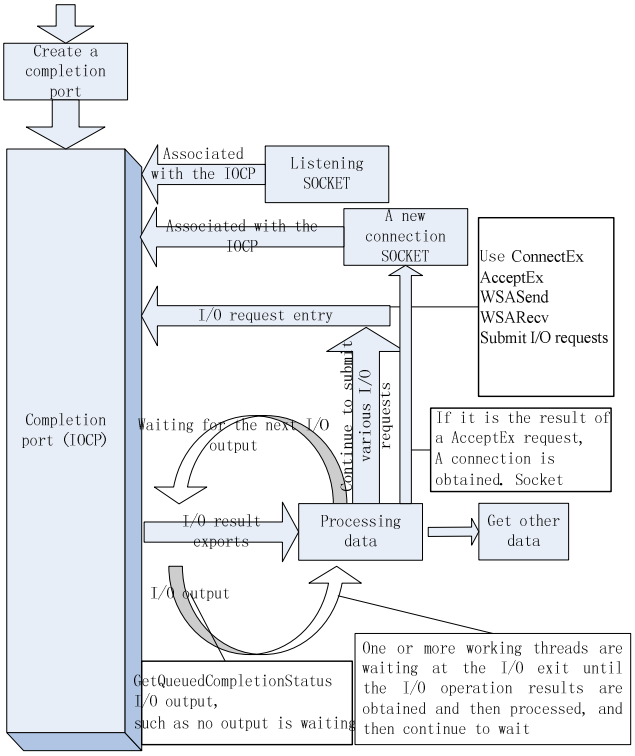
Worker threads to serve the completion port, used for processing to complete the I/O completion port notification, completion port allows the use of multi-threaded mechanism, the worker thread can create several, from the completion port management. Without I/O completion notification packets, then all threads in the completion port to wait, if I / 0 to complete the arrival notification, is received by completion port wake a worker thread I / 0 for notification, for processing. Completion port automatically work thread scheduling, wake up work which thread is determined by the completion port.

3. The operating mechanism.[2][3][4][5]

Thread is a loop operation, Windows provides a set of API enables the asynchronous blocking type operation of the network.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.66

Using a worker thread calls GetQueuedCompletionStatus, repeatedly can get when I / 0 for notification, and obtained the corresponding information of I / 0 for notification. Operation principle is as follows:



III. MEMORY ALLOCATION AND RECOVERY

A. *MEMORY CONTROL PRINCIPLE*[6][7]

In order to make the operation of the server can performance is essential to improve, we must use the memory to the IOCP necessary management. Normally when we use the memory by calling the new, delete operation to complete the application for and release, but when in high-speed mode IOCP can lead to server performance fell sharply. New and delete operation will continue this is because the user mode to kernel mode memory application and release, under a set of kernel mode memory control and memory control mechanism. This not only consumed CPU time and memory frequently change will cause the fragments of memory. So we in the allocation of resources used in recycling discarded after release, but add it to a list, when you need to use again from the list, if the list is empty, then to create a new memory space. This will be effective to enhance the overall efficiency of the system..

B. *CRITICAL AREA PROTECTION* [6][7]

The IOCP is completed by multithreading work together, so when the distribution and recovery may occur when the mutex, so here with the method of critical region lock to complete, to ensure the correctness of resources recycling and distribution. The following is a memory application pseudo code:

```
CClient * AllocateContext ()

{

Locking memory;

If (free resource nodes in the table for null)

PClient = allocate a new node space;

The else

PClient = get free resource class the first node in the table;

Remove the memory lock;

Return pClient;

}
```

C. *The safety of the resource recycling principle*

Because the system USES the asynchronous read block type, namely to read a particular Socket delivery in advance a few blocks on the completion port, when the Socket data arrived worker threads will automatically completes the data processing. In turn. So when the client out needs to be blocked in the asynchronous read process on the completion port, in turn, exit, and then arrange another time to recycle the client resources properly in order to use. In order to achieve this mechanism, we need to record every effective connection, record the number of its on the completion port congestion. So once performed as ConnectEx, accept, WSASend, WSARecv asynchronous operations function such as we are on the counter on the completion port plus one, once the exit from the completion port is minus one. So when the completion port is the counter to zero and said the current Socket without any of the asynchronous operation, then to the recycling of resources.

D. *.Tthe client disconnected resources recycling*

When the client direct call closesocket () or network anomalies are disconnected, the server then GetQueuedCompletionStatus get ERROR_NETNAME_DELETED error information. If we Socket delivery for multiple asynchronous read, cannot release the client resources directly at this moment, if the release may cause memory disorder. We adopt every exit asynchronous counter the minus one, until the counter to 0, to recycle the client resources. The code is as follows: (the worker thread)

```
UINT IOWorkerThreadProc(LPVOID pParam)
{
        ……
```

```
        BOOL bIORet =
GetQueuedCompletionStatus(hCompletionPort,&dwIoSize,(
LPDWORD) &lpClient,&lpOverlapped, INFINITE);
        if (!bIORet)  {
           DWORD dwIOError = WSAGetLastError();
           if(dwIOError != WAIT_TIMEOUT)
           {// An error occurs, but it is not a timeout event

      if(dwIOError == ERROR_NETNAME_DELETED )
                 {          ……
                    pThis-
                >m_Clients.ReleaseClient(lpClient);
                           }
                 }
          }
     ……
}
```

ReleaseClient (lpClient) automatically for the client Socket on the counter minus one, when the counter is 0 ReleaseClient Socket will be recycling into memory. Here are our in debugging the DEBUG mode, the use of the TRACE of system tracking information (including PendingIO said after our asynchronous delivery on completion port number):

```
To accept the connection socket = 2372, Client dcd9a0 =
3
Release the client's hash value mapping, socket = 2372 =
3 dcd9a0 pClient address
Disconnect the Client address = 3 dcd9a0
Close the connection = 3 dcd9a0 Client address
PClient address 3 dcd9a0 PendingIO = 3
PClient address 3 dcd9a0 PendingIO = 2
PClient address 3 dcd9a0 PendingIO = 1
PClient address 3 dcd9a0 PendingIO = 0
PClient address = 3 dcd9a0 release
```

It can be seen when closing the client thread exit will, in turn, the induction process, finally release the client resources.

E. Socket pooling and object pool allocation policy

We must reuse the socket in order to improve the efficiency of the network so that the network connection will be disconnected with speed. This is the IOCP another advantage. We can on server startup, according to our biggest service number is expected to, will all the socket resource allocation. In general each socket must correspond to a client object, used to record some of the client's information, the object pool can also and socket binding and distribution in advance. Before the service running all the large objects are pre-allocated memory resources.

F. WSAENOBUFS error problem

When we do server stress test is strange problem looks like a deadlock occurs, or memory leaks. This is actually the WSAENOBUFS error.

Along with the server every overlapped send and receive operation, data memory may be locked. When memory is locked, it cannot exceed physical memory page. Strong operating system behavior can be locked memory size set a limit. When the upper limit, overlapped operations will fail, and send the WSAENOBUFS error.

If a server in offers many overlapped receives on each connection, as the growth of the number of connections, will soon reach the limit. If the server can expect to have to deal with quite a number of concurrent client, the server can only reply on each connection received a zero byte. Has nothing to do this is because did not receive operation and memory, memory does not need to be locked. Using this method, each socket receiving memory should be complete, this is because, once a zero byte receive operation is completed, the server only for receiving memory sockets so return to a non-blocking receive data memory. Using WSAEWOULDBLOCK, when a non-blocking receive fail, no data is blocked. The purpose of this design is that at the expense of data throughput, able to handle the large number of simultaneous connections.

We have provided a simple possible solution to the WSAENOBUFS error. For 0 bytes of memory, we adopt WSARead () function. When the call is completed, we know that the data in the TCP/IP stack by adopting several asynchronous WSARead () function to read MAXIMUMPACKAGESIZE memory. This method only when data reach locking physical memory, solves the WSAENOBUFS problem.

But this scheme reduces the throughput of the server. The code is as follows:

```
void CIocpNet::OnZeroByteRead(ClientContext
*pContext,CIocpNetBuffer *pOverlapBuff)
{
        if(pContext)
        {DWORD dwIoSize=0;
        ULONG ulFlags = MSG_PARTIAL;
     ……
     }
        pOverlapBuff-
>SetOperation(IOZeroReadCompleted);
        pOverlapBuff->SetupZeroByteRead();
                UINT nRetVal = WSARecv(pContext-
>m_Socket,
                        pOverlapBuff->GetWSABuffer(),
                        1,
                        &dwIoSize,
                        &ulFlags,
                        &pOverlapBuff->m_ol,
                        NULL);

        ……
        }
}
```

G. *Packet scheduling problems*

Also discussed the issue in refs. Although using IOCP, can make the data in the order they are sent by reliable processing, but thread table is the result of the actual work thread to complete the order is uncertain. For example, if you have two I/O worker threads, and you should receive "byte chunk 1, byte chunk 2, byte chunk 3", you can deal with them in the wrong order, namely "byte chunk 2, byte chunk 1, byte chunk 3". This also means that when you pass the request to the IO completion port to send data, the data is actually be reordered after sending.

A practical solution to the problem is that our memory for increase the order number, and processed according to the sequence number of memory. Mean, is not correct number of memory is saved backup, and for performance reasons, we will memory stored in the hash table. We adopt GetNextSendBuffer and GetNextReadBuffer custom function complete order when receiving and sending the packet.

IV. THE EXPERIMENTAL RESULTS

A. *Experimental environment*

This article mainly from the client when the high concurrency memory consumption and fast connection on the server's response time of web application performance testing and analysis.

Server configuration for dual CPU, 2.83 GHz 2 GB of memory; Two client configuration as: double CPU, 2.83 GHz 2 GB of memory, a single client launched 10000 Socket connection to simulate high concurrency of the server's response speed. Network environment for 100 Mbps Ethernet LAN.

TABLE 1 QUICK CONNECT CLIENT RESPONSE TIME COMPARISON

Analog client number	Threading model	IOCP model
50	10	10
100	20	18
150	31	29
200	42	40
300	66	60
500	115	108
700	170	155
1000	253	210

TABLE 2 CLIENT QUICK DISCONNECT MEMORY CONSUMPTION COMPARISON

Analog client number	Thread model	IOCP model
0-50-0	1012	1013
0-100-0	1514	1417
0-150-0	2011	1968
0-200-0	2519	2452
0-300-0	2689	2452
0-500-0	2882	2452
0-700-0	2998	2452
0-1000-0	3953	2452

V. CONCLUSION

IOCP queue using kernel mode provides an efficient solution to network concurrency and throughput issues. But efficient mode of operation must provide an efficient memory management, memory management mode we give the IOCP model can work more stable and fast.

REFERENCES

[1] Anthony Jones,Jim Ohlund.Network Programming for Microsoft Windows editor[M]. Beijing: Tsinghua University press,2002

[2] Zhang Jinghua, Zhang Yuming.IOCP research and application in large-scale network communication system [J]. computer and modernization, 41; (9): 46 ~ 2004.

[3] RussinoviehME,SolomonDA. Pan Aimin, translation. In-depth analysis of the Windows operating system. Fourth edition, Beijing: electronic industry press,,2007:585 一 589..

[4] Richter J,Clakr DJ .Programming Sevrer-Side Application for Microsoft Windows 2000. Microsoft Press,2000:10-35.

[5] Draft ITU-T recommendation and final draft internationa standard joint video specification[S].(IT-TRec.H.264 ISO/IEC 14496-10AVC),7th Meeting:Pattaya,Thailand,2003.

[6] Wiegand T,Sullivan G J,Bjontegaad G,et al.Overview of the H.264/AVC video coding standard[J].IEEE Tanrs Circuits Syst Video Technol.,2003,13:560-575

[7] Wu D,Pan F.Fast intermorde decision in H.264/AVC video coding[j].IEEE Transactions on Cicruits and systems for Video Technology,2005,15(7):953-985

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Bank Partitioning Based Adaptive Page Policy in Multi-Core Memory Systems

Juan Fang
College of Computer Science
Beijing University of Technology
Beijing, China
e-mail: fangjuan@bjut.edu.cn

Jiajia Lu, Min Cai
College of Computer Science
Beijing University of Technology
Beijing, China
e-mail: lu_jiajia@emails.bjut.edu.cn,
min.cai.china@bjut.edu.cn

***Abstract*—With DRAM memory systems power consumption growing and performance declining, memory system optimization is imperative. At present most studies focus on reducing power consumption or improving performance of DRAM chips, but rarely do both. In this paper, we propose the bank partitioning based adaptive page policy to optimize both performance and power efficiency of DRAM chips, while at the same time reduce the inter-thread interference. First, we use bank partitioning to isolate memory streams of different threads. Second, we put forward an adaptive page policy to dynamically allocate the optimal page policy for each bank. Experimental results show that our scheme improves the system performance by 20.4% on average (up to 55%) and reduces DRAM power by 8% on average (up to 29%) for all workloads.**

Keywords-bank partitioning; adaptive page policy; multi-core memory systems

I. INTRODUCTION

Modern DRAM systems mainly rely on spatial locality to optimize memory power efficiency and performance. However, with the increasing number of cores on a chip, memory access streams have lower spatial locality because access streams of independent threads may be interleaved at the memory controller [1]. Unified page strategy is no longer suitable for multi-core systems because of the inter-thread interference. Recent work shows that memory power consumption is close to or even greater than that of the processor, which now accounts for 19% ~ 41% of the whole system [2]. So the interference among cores severely degrades both system performance and power efficiency. Therefore, it is important and urgent to improve DRAM performance and reduce power.

At present most of the studies focus on reducing power consumption or improving performance of DRAM chips, but rarely do both. Single improvements like cask theory can't really solve the problem. Hongzhong Zheng et al. point out that page policy has a significant impact on memory power consumption and the optimal page policy is application-dependent [3]. Because the open page policy is suitable for banks whose memory access stream is intensive, which can effectively use the spatial locality to reduce the delay and operation power. The close page policy has more effects on banks whose memory requests are rare, which would make the row buffer idle immediately after the column access and greatly increase the probability of the rank entering the power down mode to reduce background power.

To solve the above described issues, in this paper, the adaptive page policy is put forward. First, we implement the bank partitioning without modification in operating systems. Bank partitioning isolates memory access streams from different threads and retains the characteristics of memory accesses that each bank receives. Second, on the basis of bank partitioning, we propose the adaptive page policy, which dynamically allocates the optimal page policy to each bank based on the received memory accesses. Experimental results show that our proposal reduces DRAM power consumption by 11.5% on average (up to 29%) and increases system performance by 39.6% on average (up to 55%) for mixed workloads.

The rest of the paper is organized as follows: section 2 provides background information of the DRAM memory systems and related work. Section 3 presents our proposal in details. Section 4 describes the evaluation methodology and section 5 shows the results. Section 6 concludes the paper.

II. BACKGROUND AND RELATED WORK

A. The DRAM System

Bank is a two-dimensional structure of rows and columns, and data in a bank can only be accessed from the bank's row buffer [4]. There are three major steps required when accessing a data element:

1) *row activate:* according to the row address, activate the target row and load the data of the target row to the row buffer. Before reading to the row buffer, the row buffer must be in the idle mode. Otherwise we should precharge the row buffer first.

2) *column access:* according to the column address, read the data from the row buffer directly.

3) *precharge:* write the row buffer's data back and make the row buffer idle,which is coupled with *row activate*[5].

If the target row has already been in the row buffer, only column access is necessary, which is called a *row hit.* Otherwise, it is called a *row conflict*. When met with row conflict, the memory controller has to first precharge the opened row, activate the target row, and then perform column access. Obviously, *row conflict* would cause much higher access latency than row hit.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.67

B. Row Buffer Page Policy

In modern DRAM systems, each bank has a row buffer, which provides temporary data storage. For each bank, at the same time there can be only one row of data stored in the row buffer. There are two classes of page policies for the row buffer: the open page policy and close page policy. The difference between the two page policies is the time to precharge. The open page policy doesn't precharge the row in the row buffer until hit conflict happens. On the contrary, close page policy precharges a bank immediately after a column access.

C. Related Work

Previous work has studied on partitioning and page policy. DRAM bank partitioning is first proposed by Mi et al. [6], who use page coloring and XOR cache mapping to reduce inter-thread interference in chip multiprocessors. It requires modification in operating systems. Muralidhara et al. elaborate on application-aware memory channel partitioning (MCP), which maps the data of applications that are likely to severely interfere with each other to different memory channels based on applications' characteristics [7]. In general, the number of threads in the system is much more than the number of channels, so some threads must be assigned to the same channel, which cannot essentially eliminate the inter-thread interference. Mingli Xie et al. propose the application aware page policy, which assigns the same page policy for each application [8]. However, since the characteristics of memory accesses that each bank receives are different, the application aware page policy can potentially reduce the utilization of some banks. Xiaowei Shen et al. discuss a row based DRAM page policy in the multi-core system [9]. But they don't take into account the inter-thread memory interference.

III. BANK PARTITIONING BASED ADAPTIVE PAGE POLICY

In this paper, we investigate new techniques to reduce inter-thread interference in DRAM chips without modification in operating systems. We use bank partitioning to eliminate the inter-thread interference. Then we propose the adaptive page policy to optimize DRAM power consumption and performance simultaneously.

A. Bank Partitioning

Memory access streams from different threads running on different cores are interleaved and interfere with each other, which results in additional bank conflicts and performance loss. If memory access streams from different cores are mapped to different banks, the inter-thread interference can be eliminated. We propose to partition memory banks among cores to isolate their access streams and eliminate inter-thread interference. We first trace the memory accesses by the corresponding thread identifier. Memory accesses from a core can only access their specified banks. In the address mapping process, the same physical address from one thread will always be mapped to the same DRAM address. Therefore we have optimized the address mapping algorithm and the uniqueness of the address map is guaranteed.

The mapping algorithm works as follows: firstly, according to the memory access's thread identifier, we can decide which rank should be mapped. Secondly, we calculate the bank based on the address mapping scheme of the memory controller.

Through the above address mapping algorithm, memory access streams from different cores are mapped to different banks and inter-thread interference can be essentially eliminated.

B. Adaptive Page Policy

We add two counters in each bank to record the number of accesses and row buffer hits of each bank respectively. Then at the beginning of each interval, we assign a preferred page policy to each bank based on its accesses and hit rate. The assignment algorithm is shown in TABLE I. We use $accesses_t$ and $hitRate_t$ as two threshold parameters.

TABLE I. ASSIGNMENT OF PAGE POLICY TO BANKS

```
if accesses_i < accesses_t then
        assign close page to the bank
else
        if hitRate_i < hitRate_t then
                assign close page to the bank
         else
                assign open page to the bank
```

In the proposed adaptive page policy, a bank can dynamically change its own page policy based on its memory access streams. For example, if memory access streams the bank received in an interval is intensive and the row hit rate is high, we can assign the open page policy to the bank. This can reduce the number of *row activate* and *precharge*, thereby optimizing the operation power and latency. In addition, if the memory accesses the bank receives are rare or row hit rate is relatively low, then the interval of two adjacent memory accesses is too long or *row conflicts* increase. There is no need to keep the row buffer in the *precharge* mode waiting for the next memory access, so we assign the close page policy to the bank. This reduces the background power and the access latency of two adjacent memory accesses from the same row of the bank. Furthermore, the close page policy increases the opportunity for the rank to enter the power down mode, which can further reduce the background power.

IV. EVALUATION METHODOLOGY

A. Simulation Setup

We use both Gem5 [10] and DRAMSim2 [11] simulators as our experimental environment to evaluate our proposed bank partitioning based adaptive page policy. We set the value of the interval to 100K memory cycles, $accesses_t$ to 200, and $hitRate_t$ to 0.5. TABLE II summarizes the simulation parameters.

TABLE II. SIMULATED SYSTEM PARAMETERS

Parameter	Configuration
Processor	4 cores, 3 GHz, out-of-order
L1 caches	32KB Inst/Data, 8-way, 64B line, LRU
L2 cache	4MB shared, 16-way, 64B line
Memory	1 channel, 2-ranks/channel, 8-banks/rank Timing: DDR3-1066 (10-10-10)

B. Evaluation Metrics

Power and system throughput are two key factors in the system. As shown in (1), DRAM power is composed of Background_Power, ACT_PRE_Power, Burst_Power and Refresh_Power.

$$\begin{aligned} \text{TOTAL\_POWER} = {} & \text{Background\_Power} \\ & + \text{ACT\_PRE\_Power} \\ & + \text{Burst\_Power} \\ & + \text{Refresh\_Power} \end{aligned} \quad (1)$$

As shown in (2), we evaluate the overall throughput of the system using Weighted Speedup as in [12]. IPC_{alone} and IPC_{shared} are the IPCs of an application when it is run alone and in a mix, respectively. The number of applications running on the system is given by N.

$$\text{WeightedSpeedup} = \sum_{i=1}^{n} \frac{IPC_i^{alone}}{IPC_i^{shared}} \quad (2)$$

C. Workloads

We use multi-programmed workloads consisting of benchmarks from the SPEC CPU2006 [13] in our experiments, as is shown in table III. We simulate each workload until one of the four threads completes 100M instructions.

TABLE III. MULTI-PROGRAMMED WORKLOADS

Workload	Benchmark
1	lbm, lbm, mcf, actusADM
2	lbm, libquantum, mcf, mcf
3	lbm, libquantum, soplex, soplex
4	lbm, libquantum, calculix, sjeng
5	lbm, soplex, perlbench, namd
6	mcf, cactusADM, gcc, perlbench
7	lbm, mcf, perlbench, sjeng
8	perlbench, perlbench, namd, sjeng
9	bzip2, perlbench, gromacs, calculix
10	Hmmer, gcc, perlbench, gromacs

In order to evaluate our experimental results, we carried out three groups of different experiments for three page policies: the open page policy, the close page policy and the proposed adaptive page policy.

V. RESULT ANALYSIS

A. Row-buffer Hit Rate

As Fig. 1 shows, we gather the average row-buffer hit rate for each workload to evaluate the effect of bank partitioning.

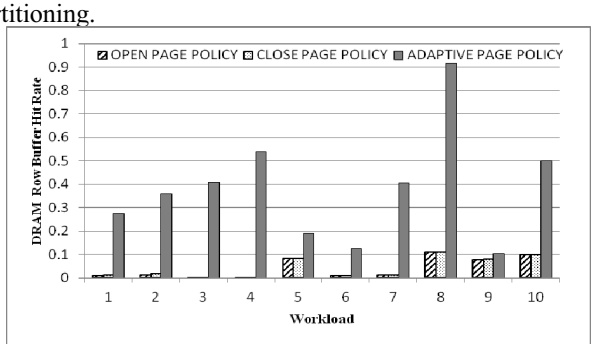


Figure 1. Average DRAM row buffer hit rate of each workload

In unified open or close page, the row buffer hit rate of all workloads are all low, especially for workload 3 and 4 (which is negligible). This is caused by inter-thread interference. In contrast, bank partitioning based adaptive page policy can greatly increase the row-buffer hit rate of each workload by isolating memory access streams from different threads.

B. System Throughput

We evaluate the system throughput using weighted speedup. The results are normalized to OPEN PAGE POLICY.

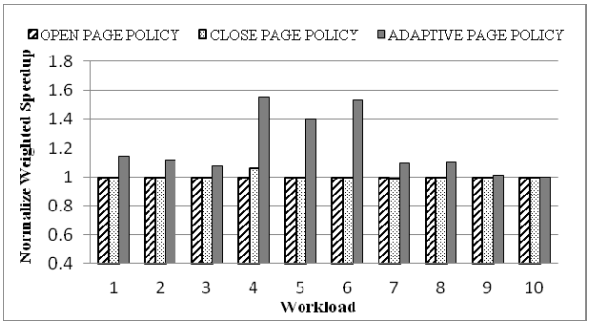


Figure 2. System throughput

Fig. 2 shows the performance impact of our proposal. For workload 1 to 3, the system performance increases by 11.2% on average (up to 14%). By reducing the number of *row activate*s and *precharge*s, the adaptive page policy reduces the latency of two adjacent memory accesses for banks whose memory accesses are intensive and row buffer hit is high. For workload 4 to 7, the effect of adaptive page policy is more obvious, which increases system performance by

39.6% on average (up to 55%). By closing the row buffer immediately after the column access, the adaptive page policy shortens the interval of memory accesses for banks whose memory accesses are rare or hit rate is low, which improves the system performance. For workload 8 to 10, the memory accesses are rare, so there is no obvious effect on system performance, which is similar to that of a unified page policy.

C. DRAM Power consumption

Fig. 3 shows the DRAM power consumption of different groups. The results are normalized to OPEN PAGE POLICY.

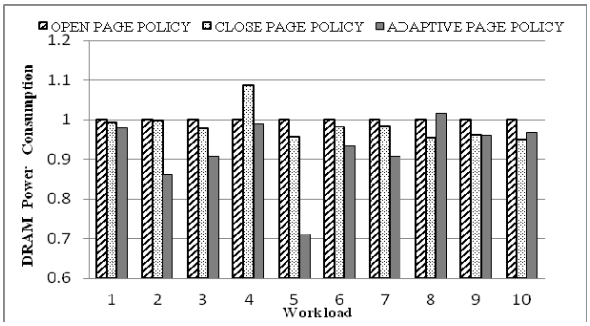


Figure 3. DRAM power consumption

For workload 1 to 3, their power consumption is mainly operation power. By reducing the number of *row activate* and *precharge*, the adaptive page policy greatly reduces DRAM power consumption by 8.4% on average (up to 14%). For workload 4 to 7, their power consumption is mainly background power. We reduce DRAM power consumption by 11.5% on average (up to 29%). By assigning the close page policy to banks whose memory accesses are rare or hit rate is low, the adaptive page policy increases the probability of the rank entering the power down mode to reduce background power. For the workload 8 to 10, memory accesses are rare. There is no obvious effect on the DRAM power which is similar to that of a unified page policy.

VI. CONCLUSION

In this paper, we present the bank partitioning based adaptive page policy to reduce the inter-thread interference among different applications and therefore optimize both system performance and power consumption. In the future, we are planning to further improve the system performance. We can dynamically change the number of banks per thread based on the demand of each thread.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China under Grant No. 61202076, No. 61202062. The authors would like to thank the reviewers for their efforts and for providing helpful suggestions which have led to several important improvements in our work. We would also like to thank all the teachers and students in our laboratory for useful discussions.

REFERENCES

[1] A. N. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. P. Jouppi, "Rethinking DRAM design and organization for energy-constrained multi-cores," ACM SIGARCH Computer Architecture News 38.3 (2010), pp. 175-186.

[2] B. Diniz, D. Guedes, J. Meira, R. Bianchini, "Limiting the power consumption of main memory," ACM SIGARCH Computer Architecture News, pp. 290-301. 35.2 (2007).

[3] Zheng, Hongzhong, and Zhichun Zhu, "Power and performance trade-offs in contemporary dram system designs for multicore processors," Computers, IEEE Transactions on 59.8 (2010), pp. 1033-1046.

[4] Ikeda, Takashi, and Kenji Kise, "Application Aware DRAM Bank Partitioning in CMP," Parallel and Distributed Systems (ICPADS), 2013 International Conference on. IEEE, 2013, pp. 349-356.

[5] Cooper-Balis, Elliott, and Bruce Jacob, "Fine-grained activation for power reduction in DRAM," IEEE Micro 3 (2010), pp. 34-47.

[6] Mi. Wei, Feng. X, Xue. J, Jia. Y, "Software-hardware cooperative DRAM bank partitioning for chip multiprocessors," Network and parallel computing. Springer Berlin Heidelberg, 2010, pp. 329-343.

[7] Muralidhara. S. P, Subramanian. L, Mutlu. O, "Reducing memory interference in multicore systems via application-aware memory channel partitioning," Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2011, pp. 374-385.

[8] Xie, Mingli, Dong Tong, Yi Feng, kan Huang, "Page policy control with memory partitioning for DRAM performance and power efficien cy," Proceedings of the International Symposium on Low Power Elect ronics and Design. IEEE Press, 2013, pp. 298-303

[9] Shen X, Song F, Meng H, et al. "RBPP: A Row Based DRAM Page Policy for the Many-core Era," 2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), Dec 2014

[10] Binkert, Nathan, et al. "The gem5 simulator," ACM SIGARCH Computer Architecture News 39.2 (2011), pp. 1-7.

[11] Rosenfeld, Paul, Elliott Cooper-Balis, and Bruce Jacob, "DRAMSim2: A cycle accurate memory system simulator," Computer Architecture Letters 10.1 (2011), pp. 16-19.

[12] Jeong, Min Kyu, "Balancing DRAM locality and parallelism in shared memory CMP systems," High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on. IEEE, 2012, pp. 1-12.

[13] Henning, John L, "SPEC CPU2006 benchmark descriptions," ACM SIGARCH Computer Architecture News 34.4 (2006), pp. 1-17.

Research and Implementation of Production Rapidly Design and Simulation Verification System Framework

Fangjing GUAN[1]
School of Internet of Things Engineering
WuXi City College Of Vocational Technology
WUXI, China
guan_fj@163.com

Haitang ZHU [2]
School of Internet of Things Engineering
WuXi City College Of Vocational Technology
WUXI, China

Zhifeng TIAN [3]
China Ship Scientific Research Center
WUXI, China

***Abstract*—The system framework applied to multi-type and multi-industry product rapid design simulation is presented. Firstly, the common problems in designing and simulation are briefly introduced, and the detailed solutions are given. Then, according to these problems, we build the construction of the system and list a detailed system framework, the main function and its realization are introduced in detail. Finally, the practicability and expansibility of the system are verified by specific application.**

Keywords- framework; designing and simulation ; Simulation verification;

I. INTRODUCATION

Complex product design is a complex process, which can accomplish data management, performance analysis, simulation and test.it relates to Structural Subjects, flow field, heat transfer, control subject. In the design simulation process, the problem is mainly reflected:

(1) Design and simulation is very professional, so engineers cost much time to skillfully master design.(2) complex operation and repeatability are also important problem.(3) normal tools is lack of industry expertise.(4) knowledge experience in design is easy to lost, knowledge is mostly distributed in the personal computer or in the minds of the engineer. (5) a large number of Software and Hardware equipments are purchaseed in a institution, but these equipments are scattered ,so it is unfavorable to manage and repair these resources (6) simulation verification is unreliable, the simulation technology is increasingly widespread. The simulation model and the effectiveness of the results are concerned. According to these problems, we build the construction of the system , the main function and its realization are introduced in detail. The practicability and expansibility of the system are verified by specific application.

II. SOLUTION

A research was made on product rapidly design and simulation verification system framework, Integrated application platform was developed to complete the centralized management of design simulation ruler, knowledge, experience, data and tools. Specific functions are as follows(Figure 1):

(1)Business process auto –fitting was provided, so non-professionals can complete Rapid design and simulation verification. (2) Through secondary development, normal tools were custom-made and optimized for user. To solve professional problems , module components are built by integrating precision calculation methods and proprietary technology.(3)In order to experts divorced from repeated work ,we realize Business Process modularization (4) Knowledge Base was formed to realize the collection of knowledge, the storage of knowledge and application of knowledge, so knowledge can be accumulated. (5) We build a distributed network service framework to realize remote resource scheduling and integrate resources in CAX software and hardware. (6) Optimization technology was imported in this system framework. The parameters of CAE model was gradually optimized by the data from the experiment, till the error of the experiment result and CAE simulated result was in a range.

III. SYSTEM FRAMEWORK

According to these problems, we build the construction of the system around tools, process, data and resource. The system framework was formed by component tool, Business Process modularization, data standardization, scheduling of remote resource. Following, the system modules will be introduced(Figure 2):

1)Tool integration (iTools): normal tools were encapsulated by component technique.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.68

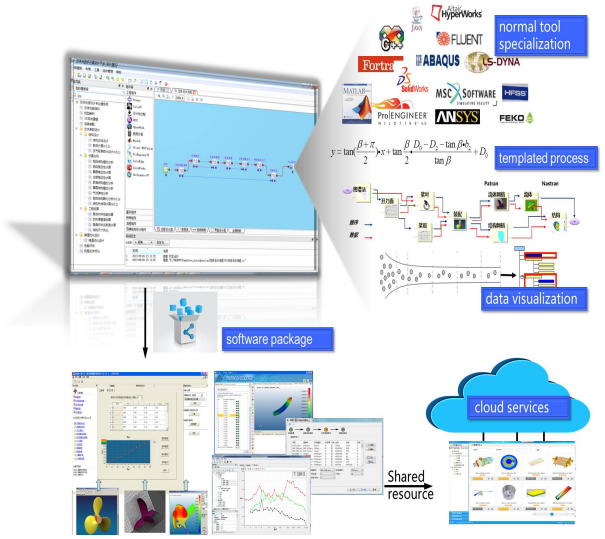


Figure1.Solution

2) Self-research Equation solver integration (iSim): it provides the professional 3D visualization environment for Self-research Equation solver.

3)Template encapsulation (iTemplates): By template technique, it encapsulated business process and cured expert experience.

4) Application Program Issuance (iApp): The man-machine interface is customized for professional template, and then it can be release to an independent application.

5)Running Control (iWorkspace): It provides neat user interface for the rapidly design of muti-solutions and rapidly simulation verification.

6) Internet Computing (iCloud): The professional template was released to a computing service which user online can call.

7) Management of data (iData): Data BOM was built through realizing Centralized management of data in simulation process.

8) Analysis and its visualization (iViewer): it has realized 2-dimension display of data, 3-dimension display of data and multi-view display of the virtual scene.

(9) Parameter Calibration of Model (iCalibrater): Model parameters are corrected automatically in methods based on numerical optimization.

IV. KEY FUNCTIONS AND IMPLEMENTATION

A. *TOOLS INTEGRATION (iTools)*

Rules and methods, which are in designing and running, are encapsulated standard components. So, engineers can rapidly complete design and simulation by the way of putting up the building blocks. Component technique is developed from Object-Oriented technology [1]. The core of component programming is to divide an application into

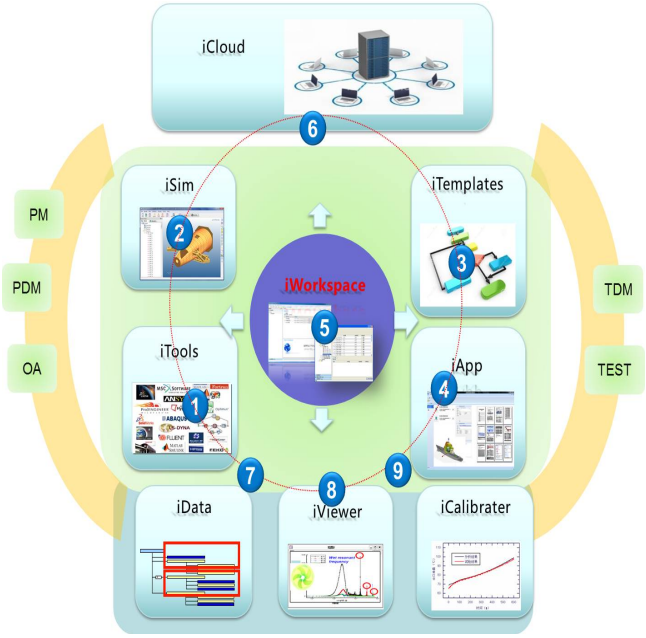


Figure2.System Framework

more components which are independent, the communications among components is realized through cooperative work. The foundation of Component application is standards. If there is no unified interface description, normative communication of components, standard Object Request and remote invocation, the component application is impossible to be realized. The current standards are primarily those developed by CORBA, EJB, COM and CLR. This integrated system is based on .NET frame, so it adopts COM standard. COM technology is platform independent, supporting object concept and realizing the separation of the decision resource description interface and implementation[2].

The system supports those tool components including ProE、NX、CATIA、Patran、Nastran、Abaqus、LS-Dyna、HyperMesh、Marc、ICEM、CFX、Fluent、ADAMS、AMESim、HFSS、FEKO、ect.

B. *TEMPLATE ENCAPSULATION (iTemplates)*

The visualization process designer on iTemplates is used to generate Business Process templates by the way of pull and push. Through template encapsulation, it gets through data flow and control flow between different tool components, supports automatic running of process and human-computer interaction running, and supports condition judgment and parallel computing.

C. *INTERNET COMPUTING SERVICE (iCloud)*

The system provides a private architecture of cloud service, realize remote resource scheduling and integrate resources in CAX software and hardware. Users can conveniently submit calculation tasks, view the running state and results by browser.

D. PARAMETER CALIBRATION OF MODEL (iCalibrater)

In normal case, Finite Element Model is simplified based on CAD model. The parameters such as model stiffness, mass and damping are different from the actual structure. There are deviations in material parameters, loading, constrained boundary, and so on. It is not possible of necessary to get the correct result, there is a deviation in analysis results and test results， even sometimes the deviation is very large. How to solve the model's validity and improve the accuracy of the model is a problem to be solved.

Through the correlation analysis method of simulation and the experimental data, we can get the uncertain parameter in the simulation model . Based on the numerical optimization method, the model parameters are automatically corrected, and the simulation results are in agreement with the experimental data(Figure 3).

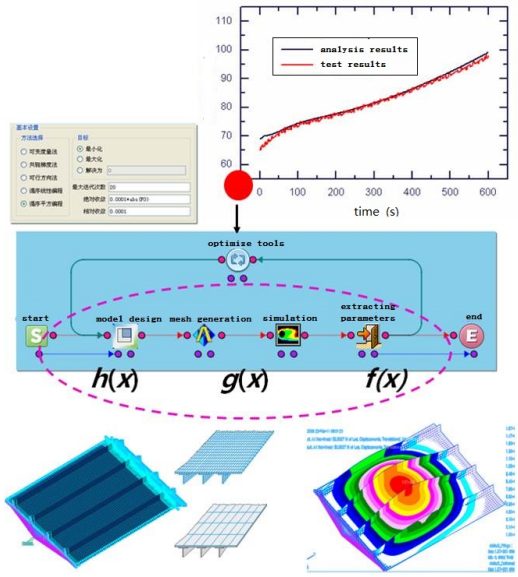


Figure3. Parameter Calibration of Model (iCalibrater)

V. APPLICATION EXAMPLES

Within the framework of the fast design and simulation verification system, we hackle the typical design and simulation process of a certain product to realize mechanical-electrical integration, the Structure and Thermal Analysis, Package of mechanism optimization design process, the Integration of ADAMS, Nastran and Matlab software, Fast iterative and comparative analysis of mult- schemem, automatic generation of reports. Finally the designer can quickly complete the simulation analysis to improve the design efficiency(Figure 4).

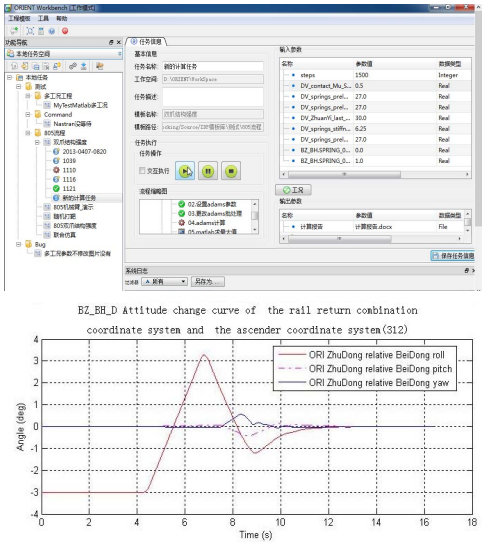


Figure4. Application example of a rapidly design and simulation verification system for an organization

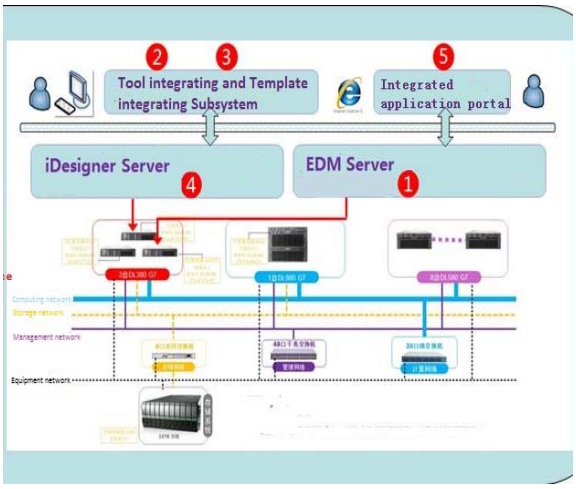


Figure5. Application example of a satellite simulation test system

We hackle satellite imaging algorithm to realize scheduling of remote high performance computing resources, image analysis of large data is realized by the user through the browser. At the same time, the data source and the satellite products are organized and managed, which greatly improve the efficiency of the satellite image processing(Figure 5).

VI. CONCLUSION

In this paper, the product design and simulation verification system has been successfully applied to many types of products. The system runs stably and reliably, and has achieved good results. We use components and template technology to improve the generality, flexibility and extensibility of the software. According to the needs of specific product design, rapid customization to develop

professional function, greatly relationship the software development cycle, saving the cost.

REFERENCES

[1] Zhang Jie, Zhang Anmin, Wang Mingbo," Design of General Test Software Platform Based on Component Technology", Ship Electronic Engineering, Vol.31 NO.7, pp.118-121.

[2] LI Hai, GUO Bin-bin, "Research on Integration Technology of HLA Based on Com Server Component", Transactions of Beijing Institute of Technology,Vol.26 NO.2, Feb 2006,pp.162-165.

[3] SHAO Li-shuo,"HLA and Component Technology Based Fleet Reconnoissance Simulation System", Computer&Network, 2012(2),pp.68-71.

[4] GUAN Fangjing, ZHU Haitang, "Design and implementation of integrated design system based on Component Technology", Computer CD Software and Applications, 2014(22).

[5] SUN Jianxun, ZHANG Liqiang,"Integrated design platform of conceptual multidiscipline_ ary for aerocrafts", Computer Integrated Manufacturing Systems, Vol.18 NO.1, Jan.2012.

[6] LAI Ming-zhu,DUAN Zhi-ming,Liu Su-yan,ZHANG Guo-yin, " Integration Research of Multi-domain Collaborative Simulation Model Base on HLA", Computer Enigneering , Vol.38 NO.16, August 2012.

[7] WANG Huai-xiao,LIU Jian-yong,LIU Ying,"HLA-based distrib_ uted visual simulation of operational entity rivalry", Journal of Computer Applications,Vol.31 Suppl.1, June 2011.

[8] WangZi,"Reasearch on the Architecture of Trajectory Module for a HLA-Based Sounding Rocket Simulation System", Chinese Academy of Sciences ,May.2010.

User Classification Method of P2P Network Based on Clustering

Shidong Zhang
State Grid Shandong Electric Power Research Institute
Jinan, China
huanazhang@163.com

Yanzhen Li
State Grid Qingdao Power Supply Company
Qingdao, China
gdiry @163.com

Zhimin Shao
State Grid Shandong Electric Power Research Institute
Jinan, China
shao_zhimin@163.com

Yong Sun
State Grid Shandong Electric Power Research Institute
Jinan, China
sunyongsd@126.com

***Abstract*—In this paper, we study the approach of user classification and service optimization for the file sharing application to P2P network. The user classification model is proposed which is based on user interest similarity and time feature similarity, and the method for determining user similarity is given. By analyzing the component features of user classification model, the results show that the model is consistent with the behavior of network users, and the efficiency and success rate of network resource search can be improved.**

Keywords: P2P Network; File Sharing Service; Clustering Analyze; User Classification

I. INTRODUCTION

In recent years, the application of peer to peer file sharing has been recognized by the Internet users, it makes full use of the idle resources of nodes in the Internet, and can also guarantee a very strong fault tolerance and ability to resist attack. For the service application, the resource search mechanism and the resource acquisition problem are the core of the network system, P2P network organization method can adapt to the highly dynamic changes of self-organization network. But it is also due to the dynamic nature of the network, the resource searching and acquisition have a lot of uncertainty, and the efficiency is very difficult to guarantee.

File sharing users always only interested in some of the resources in the P2P system[1,2], and with the change of resource popularity, user's interest also changes with time. In this paper, we analyse the clustering coefficient of P2P networks, and verify the existence of this phenomenon. The characteristics of nodes is widely used in resource search service, reconstructing the topology of P2P network according to the relationship between the nodes is an important method to improve the performance of file sharing service. The nodes of same theme are clustered together in semantic overlay network (SON)[4], according to a kind of semantic classification system, the nodes are, and form a number of SON, so as to form a number of sub SON. Node vector is introduced to describe the contents of nodes in GES model [5], the topology of the overlay is reconstructed periodically, and the semantic related nodes are organized into semantic groups. CSS[6] is an improved version of the GES model, the contents of the nodes are classified, and the virtual connection is established between the similar classes. The above research work is based on the following assumptions which is proved in literature [1,3] , if different nodes have similar content, they will issue the same query, that is, the content of the nodes on the node to represent their query behavior, so it is effective to locate the file sharing service of P2P network directly from the user interest characteristic. However, the randomness of user behavior makes the stability of the P2P network based on semantic is not guaranteed, and it's difficult to realize the semantic network by reconstructing topology. Therefore, the use of semantic information in the P2P network should be further studied.

For the research of user's online law, user query can be expressed as random sequence, common ways to describe discrete time series include AR(P) model, MA(Q) model and Markov chain. The former two models can be used for time series data fitting, its principle is to use the known model parameters and historical data to predict the value and probability of a certain moment. The Markov chain is used to describe the stochastic process with a Markov property, the conditional distribution of the future state and the independence of the past state, which depends only on the present state [8]. To the research of file sharing service, we are more concerned the total time of the process for resource acquisition than the transfer probability of a certain time period. Therefore no matter what sequence representation method is adopted, it is necessary to compare the similarity of the existing time series.

II. USER CLASSIFICATION MODEL

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

To the process of file sharing service, P2P network user activity is divided into two stages: resource search and resource download. With user interest, resource requirements can be obtained with high hit rate, while matching the online rules and their own resources can also improve the success rate of the download process. In view of the above situation, user interest analysis and user time feature analysis will be used in the selection of the users in this section, and the user classification model is expressed as:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.69

$$similarity_{All}(U_1,U_2)=(similarity_{In}(U_1,U_2),similarity_{Ti}(U_1,U_2)) \quad (1)$$

Among them, $similarity_{Ti}(U_1,U_2)$ is online matching similarity which reflects the time effectiveness of the node request resource to the users of P2P file sharing service, and $similarity_{In}(U_1,U_2)$ is interest similarity which reflects the validity of users' interest. According to the actual needs of users, $similarity_{Ti}(U_1,U_2)$ can shorten the acquisition time of shared resources, and $similarity_{In}(U_1,U_2)$ can improve the hit probability of resource requests.

III. USER INTEREST FEATURE VECTOR

Lewin is German social psychologist, and he proposed a famous human behavior equation which expresses as:

$$B=f(P*E) \quad (2)$$

In this equation, P is individual variables (genetic, personality, etc.), E is environmental variables (natural and social conditions), and F indicates their intrinsic relationship. The most intuitive performance of user interest is the number of downloads to a certain resources which are usually audio and video files, keywords of resources will be used for describing user interest in this section, and the sample of resources description is shown in finger 1. From the finger 1, we can see that the file sharing system is the requirement of fuzzy query, and the description of the resource class is more standardized, the user interest representation is a equation 3.

$$U_i=\langle K_i,W_i\rangle=\{(k_{i1},w_{i1}),(k_{i2},w_{i2}),...,(k_{in},w_{in})\} \quad (3)$$

In this equation, we use k_{i1} to represent the subject of interest to the user, and w_{i1} represent the interest weight of $User_i$ on k_{i1} that reflects the user's preferences, its value updates with the changes of user's access frequency and operation in a period of time.

```
- <FileInfo>
  - <Name>
      <Nameorder>order</Nameorder>
      <Namedata>data</Namedata>
    </Name>
    <Length>file length</Length>
    <Type>file type</Type>
  </FileInfo>
- <IMDB>
    <Grade>IMDB grade</Grade>
    <URL>IMDB grade</URL>
  </IMDB>
- <FileDes>
  - <FileDescribe>
      <BI>brief introduction of files</BI>
      <Extend>other information</Extend>
    </FileDescribe>
  - <Userdata>
      <Publisher>brief introduction of files</Publisher>
    - <Peer>
        <Peerorder>order</Peerorder>
        <Peerdata>data</Peerdata>
      </Peer>
    </Userdata>
  </FileDes>
```

Figure 1. Resource description sample

The nodes of file sharing network have many social attributes, that there is a correlation between each other[9]. At the same time, interest should have properties that can be objectively expressed, such as hierarchy and repellence, the former also known as inclusive, that is, the classification of interest, if there is coverage between nodes, and their relationship should be expressed by hierarchical structure. The latter indicates the difference of nodes that should be in quantitative terms.

This paper will document classification standards in actual system which is shown in figure 2, the tree structure is established to characterize the network resources, and then the interest of each resource is based on the resource classification tree, by which distinguish the degree of interest. The specific process is as follows: for each user access, breadth first traversal starting from the root of tree of interest, if the file belongs to a type, sets the type value to 1, and then continue to depth first traversal of the tree. By this way, a binary representation of the document in a tree structure is obtained which named *TID* (type ID). Because each file has a unique ID, the weighted sum of all the resources in one node is the interest to a certain resource classification which expressed as IST . And we will get the interest vector of the peer file sharing user U_i , by mapping *TID* and *IST* to the equation 3.

According to the network user interest vector, shared information content method is used to judge the user's interest similarity. It proposes that the similarity of two vectors depends on the degree of their shared resource information, the greater the sharing of resources between the two users, the more similar to the interest, and the sharing resource can be replaced with access ratio of similar resources to P2P network file sharing service.

$$U(PT_i)=W_i/\sum_{j=1}^{n}W_j \quad (4)$$

In the equation 4, we use k_{i1} to represent the subject of interest to the user, and w_{i1} represent the interest weight of $User_i$ on k_{i1} that reflects the user's preferences, its value updates with the changes of user's access frequency and operation in a period of time.

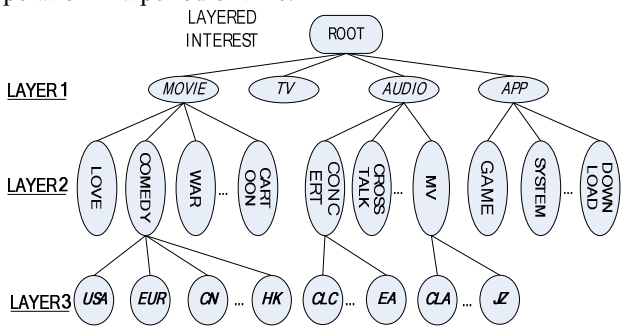


Figure 2. Resource classification standards

In the equation 3, we use PT_i to represent the probability of T_i , which is the radio of the number of the times of T_i and all resource. We only considers the statistical rules of the leaf nodes in this paper, and cosine similarity is used to the similarity between feature vectors, as shown in equation 5.

$$similarity_{In}(U_1,U_2)=\frac{\sum_{i=1}^{k}U_1(PT_i)\times U_2(PT_i)}{\sqrt{\sum_{i=1}^{k}U_1^2(PT_i)}\times\sqrt{\sum_{i=1}^{k}U_1^2(PT_i)}} \quad (5)$$

$U_1(PT_i)$ and $U_2(PT_i)$ are the representation of possibility of resource i to all resource classification, which is recorded in their own nodes, and it is open to all users.

IV. TIME FEATURE

The time characteristic of P2P file is the user's time feature, which can describe user behaviour from time dimension. Indicative function $\delta(T_i)$ is introduced to represent the state of user node, if node was online the $\delta(T_i)$ will be marked 1, otherwise 0.

$$\delta(T_i) = \begin{cases} 0, node \quad is \quad inactivity \quad in \quad T_i \\ 1, node \quad is \quad activity \quad in \quad T_i \end{cases} \tag{6}$$

File sharing network user time characteristics can be expressed as a gene sequence, which is shown in equation 7. It is possible to encounter the "Hamming cliff" [10], if we only use Hamming distance to determine the similarity between individual time series. In view of this situation, similarity operator is used to temporal similarity of different individual, and the operator is defined as follows:

$$similarity_T(U_1, U_2) = \sum_{l_i}^{l_n} \delta(T_i) \tag{7}$$

In this operator, i is the individual time series of the i locus, and the $l_1, l_2 ... l_n$ represents the same value of the two individual genes in the corresponding gene. The similarity operator is in fact a number of two different combinations of the same atoms.

V. MODEL ANALYSIS

A. User interest analysis

In order to analyze the user's interest, we need to classify the resources and pre-process. The data come from the BYR BT system, which include 3537 resource records and 200 users' information, specific instructions are as follow table 1.

Table 1. Resource classification

Category	Subclass	Number
Movie	10	1366
Tv	5	494
Music	6	130
Game	3	124
Comic	6	549
Variety	4	111
Software	4	92
Doc	11	347
Sports	4	90
Record	10	234

Figure 3 and Figure 4 show the calculation result of mutual interest similarity between 200 nodes by using equation 4. The former figure shows that the similarity between nodes is generally not high, and the interest similarity between nodes is maintained at 0.4 to 0.2 ranges, and the average value is about 0.231. The latter figure is t-test for the distribution of node interest similarity, it shows that the distribution of the nodes' interest is satisfied with the normal distribution assumption, the mean value of the fitting curve was 0.2782, and the confidence interval of the mean value is 0.95.

Resource coincidence probability is defined in this paper to verify the validity of the representation method of interest similarity, which is the percentage of total overlap of two user nodes, and its calculation method as equation 8.

$$\eta = \left(\frac{N_s}{N_{tol}^1} + \frac{N_s}{N_{tol}^2} \right) / 2 \tag{8}$$

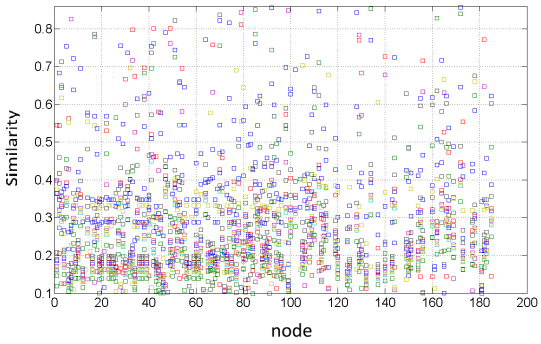


Figure 3. User interest similarity

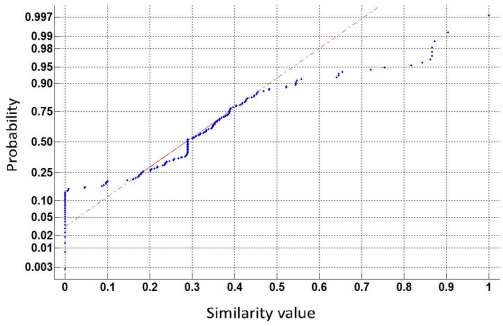


Figure 4. Verification of user interest similarity

To equation 8, N_s is the number of repeat resources for two users, N_{tol} is the total number of resources shared by the user. Comparison of the relationship between the coincidence probability of resources and the interest of the nodes is shown in figure 5. We can see that with the increase of the similarity of the user interest, the higher the rate of the users' resource duplication, which can indicate the faster return for searching to the same interest class user, and also verify the validity of the proposed method.

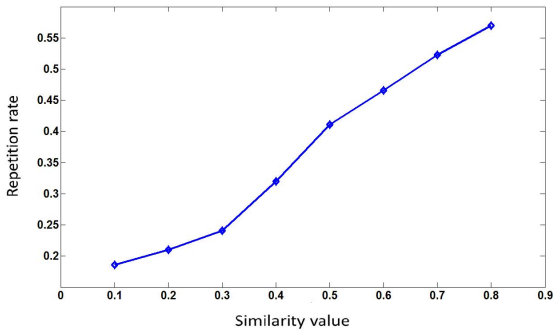


Figure 5. Interest similarity and resource repetition rate

B. *Time feature analysis*

In the actual file sharing network of BT mechanism, the relationship between time similarity and the number of copies is shown in figure 6. As can be seen from the graph, the similarity value of the access time is not very high, the mean value is 0.1903. The average number of copies was between 35 and 10, and the average value is about 11.5578. According to the theory of resource reuse, the number of copy $\hbar$, the availability of resource objects α and the availability of copy p satisfy the equation 9[11], so we can calculate the resource availability in real networks is 0.9128.

$$\hbar = \log(1-\alpha) \Big/ \log(1-p) \tag{9}$$

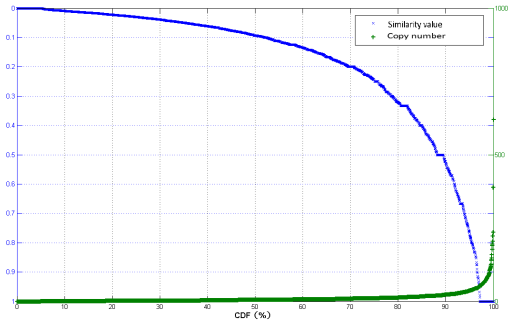


Figure 6. Time similarity and resource copy number

VI. CONCLUSION

In this paper, we presents that the users are usually interested in some of the resources, and put up distinct interest features and time characteristics to file sharing service in P2P network. User access system behavior is random, and the high clustering coefficient of user can be used to guide the user's service node selection.

REFERENCES

[1] Arturo Crespo, Hectpor Garcia Molina. Semantic overlay networks for P2P systems. AGENTS AND PEER-TO-PEER COMPUTING, 3601(2005): 1-13, 2005.

[2] Klemm Alexander, Lindemann Christoph, Vernon Mary K. Characterizing the query behavior in peer- to-peer file sharing systems. In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. Taormina,Silicy,Italy, 2004,p.55- 67

[3] A. Barrat, M. Weigt. On the properties of small-world networks. European Physical Journal, 13(3): 547–560, 2000.

[4] Kjetil , Christos Doulkeridis, Michalis Vazirgiannis. Semantic overlays for P2P web searching. AUEB Technical report, 2005.

[5] Zhu Yingwu, Hu Yiming. Enhancing search performance on gnutella- like P2P systems. IEEE Transaction on Parallel and Distributed Systems,17(12) : 1482- 1495,2006.

[6] Huang Juncheng, Li Xiuqi, Wu Jie. A class- basedsearch system in unstructured P2P networks. In Proceedings of 21st International Conference on Advanced Networking and Applications. Niagara Falls, Canada, 2007,p.76- 83.

[7] Guanglu Gong, Minping Qian. Tutorial of random process application. Tsinghua University press. 2004, Beijing.

[8] Guanglu Gong. Introduction to probability model. People's Posts and Telecommunications Press.2007, Beijing.

[9] Xiaobo Zhou, Jian Zhou, Hancheng Lu, Peilin Honh. A Layered Interest Based Topology organizing Model for Unstructured P2P. Journal of Software, 18(12):3131−3138, 2007.

[10] Xilu Zhu, Bai Wang. Niche genetic algorithm based on cluster division, Control and Decision, 25(7):1113-1116,2009.

[11] M.Frans Kaashoek, David R.karger. Koorde: A simple Degree-optimal Distributed Hash Table. Lecture Notes in Computer Science, 2735(2003), pp.98-107, 2003.

Design of Ethernet to Optical Fiber Bridge IP Core Based on SOPC

Yongjun Zhang
Institute of information photonics and optical communications
Beijing University of Posts and Telecommunications
Beijing, China
yjzhang@bupt.edu.cn

Xiangxing Kong
Institute of information photonics and optical communications
Beijing University of Posts and Telecommunications
Beijing, China
shanekong@126.com

***Abstract*—Based on the technology of system on a programmable chip (SOPC), an Ethernet to fiber bridge IP core is developed and implemented in Xilinx FPGAs. The IP core provides a way to interconnect between Ethernet port and optical port, which can be used to address the needs of remote connection and break the transmission distance restriction of cable. The logic design of IP core is done in Xilinx development tools ISE, and then encapsulated into custom IP core in EDK tools, so that the user-defined IP core can be used in embedded system. Build SOPC system with Microblaze soft-core processor, and add the IP cores into processor platform through Axi4-Stream bus interface. By testing and validating, the results show that Ethernet to optical fiber bridge IP core fulfills the purpose to transmit Ethernet data on optical fiber.**

Keywords-SOPC; Ethernet to Optical Fiber; Microblaze; Axi4_Stream; IP Core.

I. INTRODUCTION

In recent years, the development of microelectronic technique has led to sharp increase in number of logic units of a single FPGA chip, which makes it possible to implement the whole system on a single chip. The IP cores such as Processor, memory, peripherals, user-defined IP, as well as embedded software can be integrated into a single FPGA chip by SOPC technology [1]. SOPC solutions with its hardware and software co-design technique can not only reserve the flexibility of software programming scheme, but also increase the processing speed.

In Xilinx Platform Studio (XPS), Xilinx provides a plenty of IP cores, such as Microblaze soft-core processor, AXI4 bus, MAC controller, IIC controller, etc., which can be used to build the hardware of embedded system [2]. However, the provided IP cores can't cover all the cases, users has to customize their own IP core to meet specific design requirements, and XPS offers method for users to customize their own IP cores.

In order to monitor and control the board status of optical transmitter and receiver which are connected together by optical fiber, we develop a set of embedded network monitor platform, as is showed in figure1.With hardware-software co-design methods, TCP/IP protocols and user applications are implemented by software embedded in the Microblaze processor, peripherals and user logic are implemented by hardware. This helps reduce the difficulty of developing complex protocols like TCP/IP, and take full advantage of FPGA's parallel processing capacity [3] [4][5].

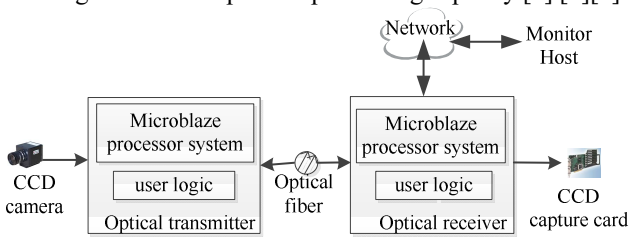


Figure 1. Diagram of embedded network monitor platform

The optical transmitter and receiver are connected by optical fiber, and the distance between them may extend to 10 kilometers. Problem arises when we need to monitor and control the status of the two optical transmitter and receiver at one side simultaneously. On other words, we need an interface converter between Ethernet port and optical fiber port, so that the monitor host can not only be connected with the local optical transceiver, but also be linked to the remote optical transceiver. In this paper, we concentrate on designing Ethernet to optical fiber bridge IP core named mac_to_aurora to fulfill the purpose to transmit Ethernet data onto optical fiber.

II. SYSTEM OVERALL DESIGN

The network monitor platform is based on an embedded system synthesized into the FPGA, and its internal logic is showed in figure 2.

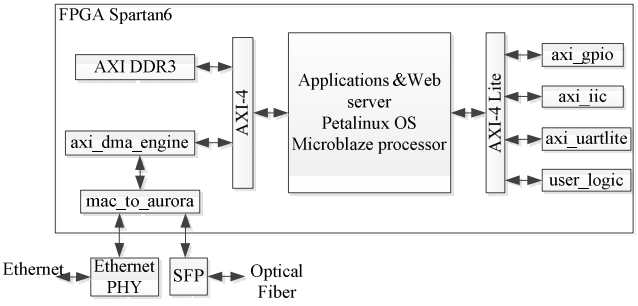


Figure 2. Internal logic of network monitor platform

In XPS, we choose Microblaze soft-core processor and add necessary peripherals IP cores to build the hardware of

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.70

the embedded system [6]. These include AXI4/AXI4-lite bus, DDR3 memory controller, user defined IP core mac_to_aurora, and other external devices connected to the FPGA. In addition, we transplant Petalinux operating system onto Microblaze processor, and develop a WEB server together with other user applications to monitor and control the status of board.

The user defined IP core mac_to_aurora used in SOPC has two functions. Firstly, it acts as MAC network controller, which is responsible for Ethernet framing protocols and error detection. Secondly, it aims to transmit Ethernet data onto Optical fiber.

III. USER DEFINED IP CORE DESIGN

A. Logic Structure of mac_to_aurora

To design IP core mac_to_aurora, we make use of the existing IP cores, like Tri-Mode Ethernet Mac IP and AURORA 8B/10B IP. Tri-Mode Ethernet Mac IP is responsible for Ethernet framing protocols and error detection, and AURORA 8B/10B IP is responsible for transforming Ethernet data into high-speed serial bit stream which is fit for transmitting on optical fiber. The logic structure of mac_to_aurora is showed in figure 3.

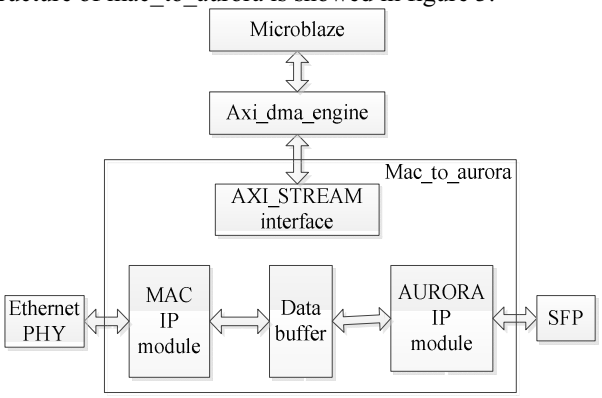


Figure 3. Structure diagram of mac_to_aurora IP core

Mac_to_aurora IP core acts as a bridge between Ethernet PHY and SFP module. The inner logic of IP core can be divided into three main sections, such as Mac IP module, data buffer module, and Aurora IP module. The interface between Ethernet PHY and Mac IP module is media independent interface (MII), which is a universal bus interface specified in the IEEE802.3 can be used to connect different types of PHY with Mac. Data buffer module stores data flooding in from Ethernet or Optical Fiber, and serves as interface adapter between Mac IP module and AURORA IP module. The user side interface of Mac IP module connects with either the AXI_STREAM bus interface or data buffer module, which can be controlled by users. The purpose is to avoid data collision, when user want to monitor the local board, user side interface of Mac IP module connects to AXI_STREAM bus interface; if monitor the remote board, then the interface connects with data buffer module.

B. Mac IP Module

To fulfill Ethernet Mac layer's function, we make use of Xilinx IP core, Tri mode Ethernet Mac, which is designed to IEEE Std 802.3-2008 specification and has comprehensive functions [7]. It supports Axi4-Stream user interface for transmit and receive frame data path. According to design requirements, we configure the IP core with following parameters. Select MII as the physical interface, the corresponding Mac speed is 10/100Mbps, choose configuration vector as the host and management interface, and disable half duplex, other options remain its default values. After the IP core is generated, we can obtain an example design under the project directory, which offers valuable examples for clock generator, physical interface, bad frame filter, timing constraints and so on and can be used to build our mac_to_aurora IP core. The use side interface of Mac IP is Axi4_Stream, which is used as a standard interface to connect components that wish to exchange data. The Axi4-Stream interface supports a wide variety of different stream types, such as transfer, packet, frame, and data stream. Here the stream type is Ethernet mac frame.

C. Aurora IP Module

Aurora 8B/10B IP core is applied to transform Ethernet data into high-speed difference serial bit stream and load it onto optical fiber. The Aurora 8B/10B protocol is an open, scalable, lightweight, link-layer protocol that can be used to move data point-to-to point access one or more high-speed serial lanes [8]. The user side interface of Aurora 8B/10B IP can directly connect to FIFO, so it will be convenient for users to customize high layer's application without paying much attention to aurora protocol internal details. Meanwhile, Aurora 8B/10B IP can provide bottom layer high-speed serial transceiver GTP's link state information, like whether the channel is set up, this helps to establish reliable communication link.

In our design, Aurora 8B/10B works in duplex mode, its user interface is Streaming, reference clock frequency is 125MHz, line width is 16bit, and line rate is 2.5Gbps. Aurora 8B/10B protocol adopts 8B/10B line code technology, which data transmission utilization ratio is 80%, so the actual line rate is 2Gbps. Ethernet data rate loaded onto optical fiber is only 100 Mbps, the bandwidth of fiber still has much redundancy, which is reserved for other use. The user interface we choose for Aurora 8B/10B IP core is Streaming interface, its timing diagram is showed in figure 4 and figure 5.

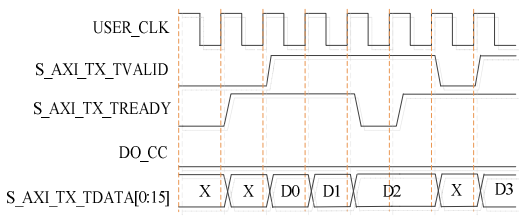


Figure 4. Typical Streaming data transfer

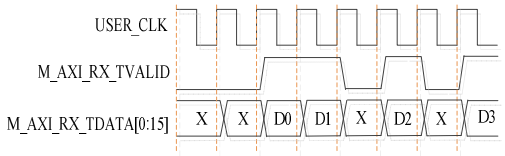


Figure 5. Typical Streaming data Reception

The streaming interface allows the Aurora 8B/10B channel to be used as a pipe [9]. When transfer data, application sends data through S_AXI_TX_TDATA port, and uses S_AXI_TX_TVALID to indicate whether the data is valid. Aurora 8B/10B IP uses S_AXI_TX_TREADY to indicate whether the channel is ready to accept data. It adopts TVALID and TREADY handshake mechanism. Only both TVALID and TREADY assert high will the data be successfully transferred. When receive data, data presents on M_AXI_RX_TDATA bus and M_AXI_RX_TVALID asserts high, if this is unacceptable, a buffer must be connected to the RX interface to hold the data until it can be used.

D. Data Buffer Module

The user side Axi4_Stream interface of Mac IP module is not compatible with Aurora IP module's user side Streaming interface, data buffer module aims to solve the problem. The incompatibility is chiefly reflected in two aspects: firstly, the clock signals of the two modules derive from different clock oscillators, clock domain crossing needs to be settled; secondly, Axi4_Stream interface of Mac IP core needs TLAST signal to denote the last byte of Mac frame, while the Streaming interface of Aurora IP is always available for transfer data after initialization, definitely, it does need a TLAST signal. In our design, we utilize asynchronous FIFO to construct data buffer. The logical circuit of transfer data buffer is showed in figure 6.

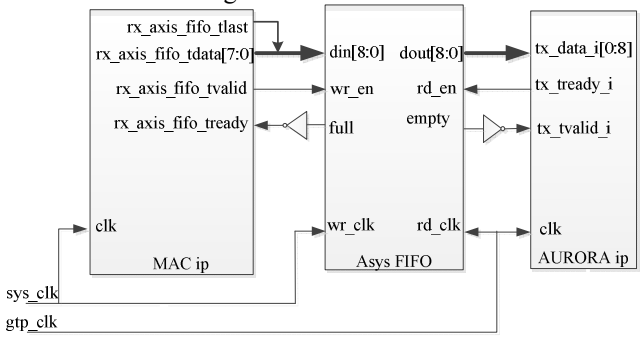


Figure 6. Logical circuit of transfer data buffer

The clock signal of Mac IP's user side interface is sys_clk, which derives from system clock source; while the clock signal of Aurora IP's user side interface derives from GTP's dedicated clock source. We combine rx_axis_fifo_tlast signal with rx_axis_fifo_tdata[7:0], and assign it to tx_data_i[0:8] as a whole. In fact, the width of tx_data_i is 16 bit, the remainder bits leave for other use. It should be noted that the byte order of rx_axis_fifo_tdata[7:0] is big endian, and that of tx_data_i[0:15] is little endian, we need to swap the order during assignment.

The reception data buffer coverts the data from Aurora IP Streaming interface into the right form adapting to Mac IP's Axi4-Stream interface. The logical circuit of reception data buffer is showed in figure 7.

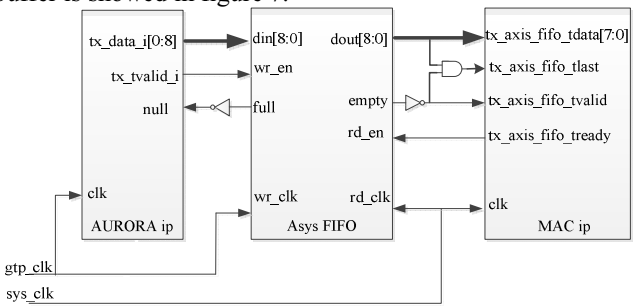


Figure 7. Logical circuit of reception data buffer

IV. IP CORE ENCAPSULATION

Until now we have finished the logic circuit of mac_to_aurora IP in ISE, before it can be utilized in our embedded system, we need to encapsulate it into custom IP core in Xilinx Platform Studio (XPS). XPS provides "create and import peripheral" (CIP) wizard which can guide users to customize their own IP cores. During the procedure, it is important to choose the suitable bus interface, which determines how the user-defined IP core will be connected to Microblaze processor. In consideration of the user side interface of Mac IP is Axi4-Stream, and that Ethernet data has the characteristics of fast rate and high burstiness, we choose Axi4-Stream as user IP's bus interface.

Unfortunately, the supported bus interface between Microblaze processor and programmable logic is limited to Axi4 and Axi4-Lite [10], IP cores with Axi4-Stream interface can't be directly connected to Microblaze processor. To connect IP core with Microblaze processor, we need to use AXI Direct Memory Access (AXI DMA) IP, which provides high-bandwidth direct memory access between the Axi4 memory mapped and Axi4-Stream IP interfaces. The mac_to_aurora IP is applied into network monitor platform.

On one hand, axi_dma_0 connects with user-defined mac_to_aurora IP through M_AXIS_MM2S and S_AXIS_S2MM interface. On the other hand, it connects to system memory MCB_DDR3 by Axi4 Memory Map Read/Write Master interface M_AXI_MM2S and M_AXI_S2MM. Besides, axi_dma_0 provides S_AXI_LITE interface to connect with Axi4-Lite bus, which provides access to axi_dma_0's initialization, status, and management registers.

After CIP wizard procedure, a template for a custom peripheral can be obtained, which contains key files to customize IP cores. To complete IP core design, we add proprietary logic to the HDL template files, and modify relevant files, such as files with '.pmd' and '.pao' as its filename suffix.

V. TEST RESULTS

The test platform is based on Xilinx Corporation's spartant-6 FPGA SP605 Evaluation Kit, on which the FPGA chip is XC6SLX45T. To validate whether user-defined mac_to_aurora IP can fulfill the function of transmitting Ethernet onto optical fiber, we make use of the following test schedule, as is showed in figure 8.

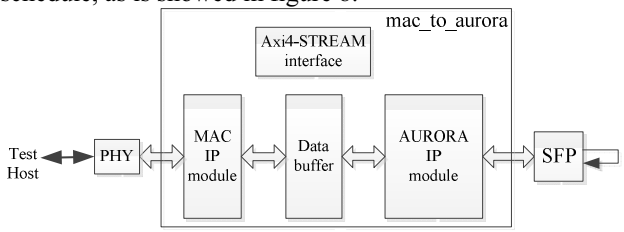


Figure 8. Test scheme

Test host keeps sending ARP packets to SP605 evaluation kit, Mac IP module accepts the packets and stores it in data buffer FIFO, and then Aurora IP module transforms Mac frame into high-speed difference serial bit stream and loads it onto optical fiber. The sent out data will loopback to mac_to_aurora IP, and finally get back to test host. During the procedure, we use Chipscope to monitor the ARP packets on-line, and compare the transfer packets with reception packets. The data waveform captured by Chipscope is showed in figure 9.

In figure 9, mii_rxd represents test host's transfer ARP packets, and mii_txd denotes the loopback reception packets; rx_axis_fifo_tdata indicates transfer data buffer's input data, and tx_axis_fifo_tdata denotes reception data buffer's output data. After magnifying the waveform, we respectively compare the transfer and reception data. Comparison results show that transfer and reception packets are perfectly the same.

VI. CONCLUSIONS

In this paper, we describe an embedded network monitor system, which is used to monitor and control boards' status. Aiming at the needs of remote connection in embedded network monitor system, a mac_to_aurora IP is designed and implemented based on the new technique of SOPC, which has the advantage of flexible structure, easy to upgrade and short development cycle. Test shows when Ethernet Mac speed is 100Mbps, the designed IP provides an interconnection between Ethernet port and optical port and fulfills the objective to transmit Ethernet data onto optical fiber in embedded system based on FPGA, the supported mac speed can be improved to 1Gbps by reconfiguring Mac IP module. This paper can provide valuable information for people searching for methods to develop custom IP core based SOPC.

REFERENCES

[1] T. S. Hall and J. O. Hamblen, "System-on-a-programmable-chip development platforms in the classroom," IEEE Trans. Educ., vol. 47, no. 4, pp. 502–507, Nov. 2004.

[2] Fu Y, Deng C, Liu X. The design of motion compensation IP core based on SOPC[C]//Intelligent Control and Information Processing (ICICIP), 2010 International Conference on. IEEE, 2010: 453-457.

[3] H.-C. Huang and C.-C. Tsai, "FPGA implementation of an embedded robust adaptive controller for autonomous omnidirectional mobileplatform," IEEE Trans. Ind. Electron., vol. 56, no. 5, pp. 1604–1616, May 2009.

[4] ZHANG Cheng, CAI Hui, CAI Hui-zhi. Gigabit Ethernet Design Based on SOPC [J]. Microelectronics & Computer, 2009, 26(2)-4

[5] De Souza R N, Muniz D N, Fidalgo A V S. Ethernet communication platform for synthesized devices in Xilinx FPGA[C]//EUROCON-International Conference on Computer as a Tool (EUROCON), 2011 IEEE. IEEE, 2011: 1-4.

[6] "EDK Concepts, Tools and Techniques: A Hands-On Guide to Effective Embedded System Design," Xilinx Inc., UG683, Version14.1, April 2012.

[7] Xilinx, (2012) LogiCORE IP Tri-Mode Ethernet MAC v5.4. Available:http://china.xilinx.com/support/documentation/ip_documentation/tri_mode_eth_mac/v5_4/pg051-tri-mode-eth-mac.pdf 350.

[8] .Xilinx, (2014) Aurora 8B/10B Protocol Specification v2.3. Available:http://china.xilinx.com/support/documentation/ip_documentation/aurora_8b10b_protocol_spec_sp002.pdf.

[9] Xilinx, (2012) LogiCORE IP Aurora 8B/10B v8.1. Available:http://china.xilinx.com/support/documentation/ip_documentation/aurora_8b10b/v8_1/aurora_8b10b_ug766.pdf

[10] Xilinx, (2013) MicroBlaze Processor Reference Guide v14.7. Aailable:http://www.xilinx.com/support/documentation/sw_manuals/xilinx14_7/mb_ref_guide.pdf

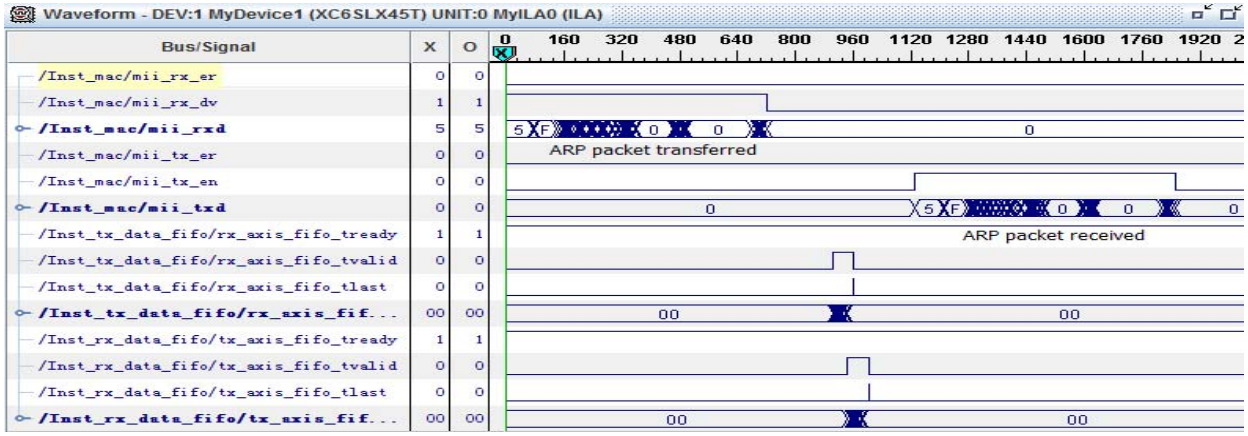


Figure 9. Data waveform captured by Chipscope

K-means Clustering Algorithm for Large-scale Chinese Commodity Information Web Based on Hadoop

Geng Yushui
School of Information
Qilu University of Technology
Jinan250353, China
gys@spu.edu.cn

Zhang Lishuo
School of Information
Qilu University of Technology
Jinan250353, China
631901036@163.com

***Abstract*—With the growing popularity of the network, product information filled in the many pages of the Internet, which you want to get the information you need on these pages tend to consider clustering information, and the current explosive growth of data so that the information mass storage condition occurs, clustering to facing the problems such as large calculation complexity and time consuming, then the traditional K-Means clustering algorithm does not meet the needs of large data environments today, so this article combined with the advantages of the Hadoop platform and MapReduce programming model is proposed the K-Means clustering algorithm for large-scale chinese commodity information Web based on Hadoop. Map function calculates the distance from the cluster center for each sample and mark to their category, Reduce function intermediate results are summarized and calculated new clustering center for the next round of iteration. Experimental results show that this method can better improve the clustering processing speed.**

Keywords-K-Means clustering algorithm; Hadoop platform; MapReduce;Cloud computing; Big Data

I. INTRODUCTION

A. Background and Significance

Currently, the network has become the primary platform for people to obtain information, the same information through the network to understand the commodity has become the most popular channels, and the information for a commodity often scattered in various web pages, product attribute information described in each page again different, you want to find the information you need product attributes from these pages should be considered for these pages are clustered in the center of the content of each page is described on the basis of its clusters, according to the clustering results easier to view the corresponding commodity attribute information. But with the development of networks, data networks explosive growth, the traditional clustering algorithms have been unable to meet the needs of large-scale data clustering, it often faced with the complexity and time-consuming calculation of such problems, this time on need to traditional clustering and cloud computing combined.

Cloud computing as a new business model has been widespread concern [1-4]. Hadoop is an easier to develop and parallel processing of large data cloud computing platform, which is mainly a strong expansion capability, low cost, high efficiency and good reliability, and other characteristics, Hadoop platform mainly Hadoop Distributed File System (HDFS) [5] and MapReduce [6] computing model in two parts, the use of Hadoop platform makes it easy to build a high-quality and efficient distribution system. These features use Hadoop platform, the traditional clustering algorithm combined with the Hadoop platform, computational complexity and time-consuming to resolve such problems as traditional clustering algorithms for large-scale web clustering encountered.

B. Related Work

There is a widespread large-scale clustering analysis and distribution of data is difficult to deal with, difficult to determine the parameters, inefficiency and poor clustering quality and other issues. There are already a number of scholars have studied the clustering algorithm parallelization and distributed execution strategies. Literature [7] is based on MPI conducted parallel K-Means clustering algorithm study and application in the process of resume data processing; Literature [8] designed cloud data mining based on Hadoop decision tree SPRINT algorithm implementation; Literature [9] proposed a distributed based on K-Means algorithm, each iteration to be transferred between nodes, a large number of data objects, resulting in large communication cost, and it is far greater than the computational cost, lower overall efficiency; Literature [10] using a cluster center broadcasts to communicate each iteration requires inter-node communication, traffic considerable; Literature [11] tried to reduce the number of communications, some of the data taken parallel idea, the relative reduction of the computational complexity of the local, but data still needs to be read many times, the cost of communication is also quite large. However, these methods are not suitable for implementing large-scale Chinese commodity information for the clustering of the page, so the need to make K-Means clustering algorithm for Chinese website features product information based on Hadoop.

C. The Paper Work

Based on the massive Chinese commodity information page features proposed K-means clustering algorithm for large-scale Chinese commodity information Web based on Hadoop, first this paper through the range of Unicode characters \u4e00-\u9fa5 extract all of the characters on the page, and then use the Chinese Academy of Sciences word software ICTCLAS [12] carried out the Chinese word to get

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.71

each word of the document, and then through stop list filtering stop words (Stopwords) , then we calculate the entire document in a word for the right of each document weight, by setting the threshold keyword extracting a collection of documents to generate a feature vector, then Kmeans algorithm iteration until convergence, the last class to mark the page belongs. In this paper, this whole process is applied on Hadoop platform through MapReduce programming framework to achieve, Using graphs on HDFS the characteristics of parallel computing, efficient processing of a large number of Chinese Web page clustering, and through experiments to verify the proposed method has better speedup and scalability.

II. DESIGN K-MEANS CLUSTERING ALGORITHM FOR LARGE-SCALE CHINESE COMMODITY INFORMATION WEB BASED ON HADOOP

MapReduce processing data set will break down the data set into many small data sets, and each small data sets can be fully parallel processing [13,14]. For K-means clustering algorithm, where each element to calculate the distance of the center of mass is to be operated independently, there is no element in the operation process interlinked, so you can use the MapReduce model to achieve parallelism. MapReduce to parallelize thinking K-means clustering algorithm is to put each iteration serial K-means clustering algorithm is converted to a MapReduce computation, and it can operate independently in parallel computing to achieve, including the distance from the center of the sample and cluster computing and local new clustering center calculation process algorithm includes pages with text vector representation of iterative K-Means algorithm, these two processes are used in the Map function, Combine function and Reduce functions to achieve.

A. Expressed MapReduce-based Web Text Vector

Due to the content of the page is the computer does not recognize the unstructured text, you want to be the clustering of these unstructured text will be converted to text vector. Web page text vector calculation, we must first calculate the value of these texts TFIDF words, and then expressed as a corresponding vector. As can be seen from the formula of TFIDF, TFIDF on MapReduce is suitable for distributed computing, word frequency (TF) only has a relationship with its total number of words in the document and how many times it appears in this document, so you can pass data partitioning, The statistics of the document word frequency IF in parallel, it can speed up the computation, word frequency obtained after calculating weights TFIDF word contains the number of the document depends on this word. So long as it can determine the number of documents that contain this word can be achieved in parallel computing TFIDF solving. And because many of the words on the page, all whichever weight to calculate the dimension text vector only larger, but also not very good exhibit center of the page, so when we calculate the value of taking a large portion of TFIDF term weight to calculate Web page text vector.

B. Units

The basic idea of parallelism based on k-means clustering algorithm MapReduce programming framework: During the execution of the serial algorithm each iteration starts the appropriate time MapReduce computing (including Map function, Combine Function and Reduce functions), to complete the data object calculate the distance cluster center and the new cluster centers.

Randomly selected from a data object k data samples as the initial focal point and stored in a file on HDFS, as a global variable is updated each iteration executed. MapReduce computation model based on process data model, data objects be treated pretreatment row fragmentation, and no correlation between slices.

1) Parallel K-Means Algorithm Map Function

Task Map function is read line by line from the HDFS file data samples to (key/value) key-value pairs, calculate each sample to the center distance, and according to the rules of the smallest distance belong to categories and tag the sample data of its clustering categories. Enter the Map function is to be clustering all of the data samples and the last round of the iterative calculation of cluster centers obtained (or initial cluster centers), data records <key, value> value pairs in each record <line number, rows> composition; the output of intermediate results <key, value> value pairs in <cluster category ID, records attribute vector> components.

Map function pseudo-code:

```
void map(<key,value>)
{ dist_min =Math.max( ); // The shortest distance is defined variables dist_min, initialize the maximum.
ID=-1;
for(i=0;i<=k-1;i++){
dist=dist (array,cluster[i]); // Define an array array, each dimension Sample values were calculated with the i-th sample cluster center distance of each dimension. if (dist<dist_min)
{ dist_min=dist;
ID=I;}
}
key=ID;
value=array; // Constructs a string that contains the coordinates of each dimension vector samples, and given the string value.
output<key,value>; // Intermediate results.
}
```

2) Parallel K-Means Algorithm Combine Function

Combine function task is to a large number of intermediate results Map function output <key, value> local statute, the key to reducing bandwidth consumption and data transmission between nodes. Enter the <key, value> value pairs corresponding <cluster category ID, records attribute vector>, for the same ID, parse out each dimension coordinate values for each record, respectively accumulate dimensional coordinates of each cluster is worth to the sum of accumulate locally, and the number of records and statistics the same ID sample; the output of <key, value>

value pairs, key for ID, value for the total number of sample records and the sum total of local clustering in each dimension.

Combine function pseudo-code:

void combine(<key,value>)

{ num=0; // Define the variable num, the number of samples and statistics.

while(value.Next( ))

{ num++; // Same statistical clustering (ID) number of samples.

for (i=1;i<=dim;i++) // dim dimension representative sample of each record.

{sum[i]+=point[i];} // Define an array sum [i], each component is initialized to 0-dimensional coordinate values were recorded for each accumulated value.

}

value=num+sum; // The total number of records stored in the string structure and the cumulative value of each.

Output<key,value>;

}

3) Reduce Function Design

Reduce function task is summarized local clustering Combine function generates results, calculate a new cluster centers for the next round of Map-Reduce function uses. Reduce the input function (input: <ID, the number of samples with partial-dimensional coordinate values accumulated value>) to obtain the mean total number of global and all-dimensional coordinate value of the accumulated value of a statistical sample, and then divided by the total number of accumulated value vector (new cluster center coordinates); output <cluster category ID, mean vector>.

Reduce the function of pseudo-code:

void Reduce(<key,value>)

{ num1=0; // Num1 variable definition, the initial value is 0, the statistical sample of the total number of the same ID.

while(value.Next( ))

{ num1=num1+num; // Statistics partial sample number (num) of the total number.

for (i=1;i<=dim;i++) //dim representative sample dimension of each record.

{sum1[i]+=point[i];} // Define an array sum1 [i], each component is initialized to 0, respectively, the sum of the partial record of each dimension of the coordinate values of the accumulated value.

}

for (i=1;i<=dim;i++) // dim dimension representative sample of each record.

mean[i]=sum1[i]/num1; // Define an array mean[i], each component is initialized to 0, the mean of each dimension record coordinates.

value=mean[i]; // Construct a string that contains the new coordinates of the center point of each dimension (vector mean) of information, given the string value.

Output<key,value>;

}

Reduce function is based on the output to get the coordinates of the new cluster centers, and update relevant documents HDFS. Calculation results consecutive rounds squared error criterion function, if the difference is less than a given threshold, then the clustering criterion function converges, the algorithm ends; Otherwise, the new cluster center to replace the last round of the central file, enter a new round of Map-Reduce computing.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Environment with Experimental Data

Experimental environment using Hadoop platform, a total of four nodes, a service node is NameNode and JobTracker, three are DataNode and TaskTracker service nodes, each node using Centos6.5 system, JDK version using jdk-7u75, Hadoop platform is hadoop -1.1.2 version.

Experimental data crawling site major commodity information, to obtain a large number of URLs, which was divided into three groups of test data set, the data set contains some number of URLs are 100,1000,5000, called datasets data1, data2, data3 .

In the experiment, because it is based on the Chinese web page clustering information to the information contained in the web tend to be right, its workload will be great, so this was chosen when the number of samples in the data set is less compared to some. Experiments in order to avoid errors during the initial cluster centers randomly selected k resulting in the experimental group when no samples were carried out 15 times to repeat the operation, and then take the average as the experimental results, the experiments were standalone processing speed, parallel processing acceleration and scalability than the index as a judge.

B. Experiment and Analysis

Select one of the nodes in the cluster as a single set of data processing environment, the selection of the three data sets clustering operation, the operation is also through text preprocessing, TFIDF calculate weights, K-means clustering algorithm.

The selection of the three data sets in Hadoop platform to build a good operation, Fig.1 is a screenshot of one of the data sets during the experiment, and the results recorded and compared with clustering result processing data when in stand-alone environment .

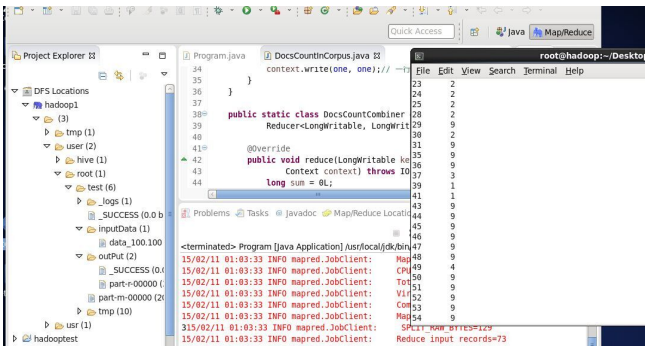


Figure 1. Hadoop platform for parallel processing data sets Screenshot.

Table shows the three data sets is a stand-alone platform with Hadoop running time comparisons.

TABLE I. STAND-ALONE ENVIRONMENT COMPARED WITH THE RESULTS OF A PARALLEL ENVIRONMENT

Dataset	Data files /KB	Number of samples	Parallel environment run time /S	Stand-alone environment running time /S
Data1	880	100	63.33	301.25
Data2	8521	1000	592.21	2610.30
Data3	39520	5000	2415.40	86512.21

From the results can be seen in Tab.1, the clustering operation processing pages in a parallel environment greatly reduced than in a stand-alone environment, handling time, Hadoop environment can improve the efficiency of a good Chinese web clustering. This article also participate in the experiment different cluster nodes for statistical data collection is completed to the same time of the experiment, and then calculate the speedup of the experiment, comparing the number of nodes in a parallel environment under the influence of different experimental efficiency (Fig.2).

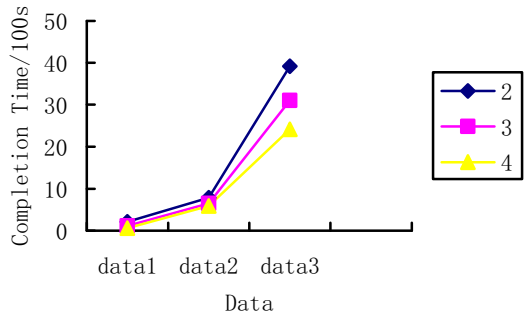


Figure 2. Chinese web clustering run parallel environment speedup results contrast.

As can be seen from Figure 2, the same set of sample data, the more nodes, the less time to perform, and so you want to improve the operating efficiency can consider increasing the number of nodes in the cluster to run.

IV. SUMMARY

Experimental results can be seen, Hadoop parallel environment was significantly higher than the efficiency of a stand-alone environment, the number of nodes also affects the efficiency of clustering, so large-scale clustering of Chinese web pages, run in the cloud platform can improve efficiency. However, since the initial cluster centers K-means algorithm is set at random, and therefore on the accuracy of the algorithm is to be improved.

ACKNOWLEDGEMENT

[1] Project supported by the projects of Shandong Province Higher Educational Science and Technology Program,China (No. J12LN20).

[2] Project supported by the projects of Shandong Province Science and Technology Development Plan,China (No. 2014GGX101052).

[3] Project supported by the projects of Shandong special independent innovation and achievements transformation,China (No. 2014ZZCX03408).

[4] Project supported by the projects of SShandong province natural science foundation,China (No. ZR2014FQ021).

REFERENCES

[5] Kryszkiewicz M.Rules in incomplete information systems [J].Information Sciences,vol. 113, 1999, pp. 271-292.

[6] Mingers J. An empirical comparison of selection measures for decision-tree induction [J].Machine Learning,vol.3, 1989, pp. 319-342.

[7] Safavian S R,Landgrebe D.A Survey of Decision Tree Classifier Methodology[R].47907.School of Electrical Engineering,Purdue University,1991, pp. 1-58.

[8] Feng Shao- rong. Research and Improvement of Decision Trees Algorithm [J]. Journal of Xiamen University(Natural Science) ,vol. 20,2007,pp. 498-500.

[9] Quinlan J R.C4.5:Programs for Machine Learbing[S].Morgan Kaufman,1993.

[10] Qian Yunhua,Liang Jiye.A new method for measuring the uncertainty in incomplete information systems[J].International Journal of Uncertainty.Fuzziness and Knowledge-Based System,2008.

[11] Feng Lina, K-Means Clustering Method and Its Application in The Resume Data in Parallel [D]. Yunnan: Yunnan University,2010.

[12] Yang Chenzhu,Data Mining Based on HADOOP [D].Chongqing: Chongqing University,2010.

[13] Kantabutra S,Couch A L.Parallel K-Means Clustering Algorithm on NOWS[J]. Technical Journal,vol. 6, 2000, pp. 243-247.

[14] Forman G,Zhang B .Distributed Data Clustering can be Efficient and Exact[J].SIGKDD Explorations,vol. 2, 2000,pp. 34-38.

[15] Boutsinas B,Gnardellis T. On Distributing the Clustering Process[J].Patter Recognition Letters,vol. 23, 2002,pp. 999-1008.

[16] ICTCLAS Project Team. ICTCLAS Chinese word segmentation system [EB/OL]. (2009-11-21). http://ictclas.org/.

[17] Liu Peng, Combat Hadoop - Shortcut to Open Access to the Cloud [M] . Beijing: Electronic Industry Press,2011,pp. 60-74.

[18] Srirama S N,Jakovits P,Vainikko E.Adapting scientific computing problems to clouds using MapReduce [J].Future Generations Computer Systems,vol. 28, 2012,pp. 184-192.

The Research on Individual Adaptive English Studying of Network Education Platform based Big Data Technology

Yanli Song
The School of Information
Guizhou University of Finance & Economics
Guiyang City,Guizhou
30585836@qq.com

***Abstract*—It appears that the adaptive English material can enhance the students' English ability. This paper model the adaptive network English education platform based big data technology and the actuality of Guizhou university students, this platform can afford the most appropriate English material, so it can advance the English level of students and reduce the differences.**

Keywords-adaptive; network education platform; big data

I. INTRODUCTION

The university English course was named as the first of the most unpopular subjects in Guizhou according to the investigation, so the passing rate of CET4 is lower than that of all national undergraduate universities because most Guizhou students don't study English. In order to change this difference, we not only need the help of network education platform for sharing the excellent education resources of the whole world, but also we should consider the adaptability of network education platform. The network education in the era of big data can record, track, master and visualize the different learning characteristics, learning needs, learning basis and learning behaviors, it can establish the learning model for different students and create the personalized learning path for different types of students. It will present the dynamic presentation according to the students' individualized learning trajectory so the content of every student is no longer stereotyped. The students will have happiness because the studying analysis based the big data make the education matched the principle of teaching according to the abilities, adapts to the change of study and the change of individuation and humanization. Therefore, the main purpose of this paper is to model the adaptive network English education platform based big data technology to raise the students' studying effects by making the students have the fundamental, independent and decisive position in the activities of the network learning, and then push the education development of Guizhou.

II. THE BIG DATA

The society is the information and the digital today, the data is full of the whole world because the development of the internet and internet of things and the cloud computing, the data also becomes the kind of new natural resources at the same time that people can exploit reasonably and effectively, it can bring more benefits and values to the life and work. Under these circumstances, the amount of data is not only increased in exponential form, but also the structure of data is becoming more and more complicated, so the big data has the new connotation than the data.

A. The Characteristics of the big data

According to searching of big data by the high technology enterprises and universities and etc, the main characteristics are as followings:

1) High density: The research is the great deal of data rated some things not depend on random sample, the value of the unit data is very low, but the analysis of the data may create value after storing and combing other data for a long time.

2) Diversity: The data that participate in the processing and analyzing can contain text, images, audio, video, database records and other data formats, and may come from multiple data sources.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.72

3) Reusable: The collected data can't depreciate because of multifarious use. Contrarily the data that combine from different sources will create great value in a certain time.

*4) The fast speeds:*The features of the big data can satisfy the real-time demands.

5) The important relevance: The big data can predict the direction of development by paying attention to the relationship between the big data and the phenomenon of things and their related phenomena.

B. The key technology of the big data

The key technology of the big data contains three fields that include the data extraction and integration, the data analysis and the data explain.

- The data extraction and integration. The source of the big data is very extensive and the data type is extremely complex, so it calls for the data extraction and integration in the data source and the unified definition to store these data. It needs to clean the data after integrated data extraction to ensure the data's quality. The existing data extraction and integration methods are mainly based on the ETL engine and the searching engine.
- The data analysis. The data analysis is the important process of big data that plays the core value. The analyses of the traditional technologies are data mining, machine learning, statistical analysis and etc. The conclusion can be used to recommend system, expert system, business intelligence and decision support system and etc.
- The data interpretation. The traditional method of explanation is only the visual display of text, charts and other computer terminal. The data interpretation ability can be introduced into the tag cloud visualization technology or human-computer interaction technology to solve. In this way it will achieve the best effect through guiding the user to enter the analysis process in the interactive process.

C. The big data processing

The source of the big data is very wide from the characteristic and the field of big data, the resulting data types and application methods differ in thousands of ways therefore. But in general, the basic process of big data is divided into four stages including the data acquisition, the data processing and integration, the data analysis and the data interpretation.

III. THE INVESTIGATION AND ANALYSIS OF ENGLISH ADAPTIVE LEARNING IN GUIZHOU UNIVERSITIES

A. The investigation

The writer has the investigation for half a year with the aim to thoroughly improve the English ability of Guizhou students. The writer analyzed the factors that influence of the students' English level. The research objects are five Guizhou universities including the Guizhou University of Finance and Economics, Guizhou Normal University, Zunyi Medical College, Bijie University and Tongren University. The total questionnaires were 7230, 6009 questionnaires were collected (the 272 of teachers and 5737 of students), the valid questionnaires were 5763.The questionnaires included the English teaching mode, teaching methods, the proportion of the teachers and students, network multimedia skills and etc. The student side related to learning strategies, learning motivation and learning situation. The aim of this paper is mainly to solve how to build the platform to improve students' English ability, so the specific content and the answers of the teachers and students are not stated.

B. The analysis of students

The adaptive learning is exist in the students, but the students' autonomy and initiative are hindered because it isn't universal. The questionnaires reflecting the reasons for low English level of the students are as followings: the autonomy of the boys is lower than girls; teachers' teaching mode and teaching method are relatively simple and old-fashioned; the shortage of teachers; the teachers graduated from non-normal specialties lack the theoretical knowledge of pedagogy, psychology, teaching materials and methods, so they can't consider the individual needs of students because of the heavy task. 85% of the students think the learning English materials can't adapt to their learning needs because it is usually too old, too rigid and too boring. The English level is too low because most students are from rural

areas, about 75% of the students didn't meet the requirements of syllabus. The students can't keep up the English learning so it causes the lack of confidence and the sense of inferiority learning in the future because they lack the language knowledge and language skills.

IV. THE CONSTRUCTION OF ADAPTIVE NETWORK ENGLISH EDUCATION PLATFORM OF GUIZHOU UNIVIERSITIES

The writer tries to design the adaptive network English education platform based on big data focusing on the students' psychological problems, let them overcome their own obstacles to embody the individuation and allow them overcome their obstacles in order to improve their English abilities. The platform can take the student as the individual in a more personal situation and provide students with more favorable condition, the students interact with adaptive learning platform and gain abilities through their original knowledge experience.

A. The application strategies of the big data in the English education platform

The individual needs of students will often change, using the big data technology can recommend dynamic page and provide personalized information service according to mining the large amount of structured and unstructured data to find the students' favorites and requirements.

Getting the deepest and valuable information depend on the management and operation of the platform, it uses the big data technology through the analysis of students' services and students' behaviors. In the learning process, different application project asks for different demands of the big data analysis processing that include the performance of data processing, the amount of data, computing speed, accuracy, real-time and diversity. The application strategies should pay attention to the followings:

- The analysis should be combined with the characteristics of different objects and process, and use the appropriate data analysis and system resource allocation strategy.
- The analysis process of the big data should be closely combined with the personal needs of the students, it can recommend the accurate and real-time personalized service according to the scene on the basis of the students' individual needs and behavior analysis.
- The analysis big data should be based on the analysis of features of the object to realize visual performance. The analysis reflects the multiple attributes and variables of individual reading activities and services.

B. The logical structure of platform

According to the above, the platform must be able to provide the students with their English materials in order to really improve the English abilities, the platform design is shown in figure 1.

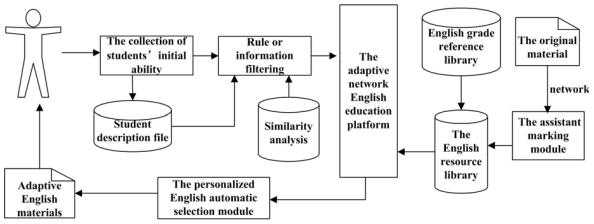


Figure 1. The logical structure of platform

C. The detailed designing of the platform

1) Mining the students' personal information

This process is the deep mining the information of students that include the procedure from students' initial ability collection to filtering based on rules or informations. The effect of students' personality information on learning is large especially for Guizhou students because their learning basis and learning interestes are different and the poor student has the heavy inferiority.

a) The student information collection and student description file

The methods of collecting student information can be divided into two types: The first one is the explicit description of the students by providing personal interest in the list or questionnaires to the student to get the students' interest. The second kind is the students' implicit description that gains the interest of students by tracking the students' behaviors because the students' behaviors can express their interests.

The idea of this paper is using the explicit description of the students when the students initial log, it can be

described by the students' implicit description with the deepening of the learning to collect the accurate and informative information as much as possible. The collected information from students are not same according to the student's individual learning, the student information constitutes the description file that is the important reference for information filtering.

b) Analyzing the students' data

Analyzing the students' data mainly adopts the method that is based on the rule that is used to realize software design rules for the individuation by providing the visual editing environment to software developers, the teaching plan is based on the application of the rule method: if the students learn the course x then recommended them learn course y. For example, if the students have finished the first of English course according to the rules then you can recommend them learn the second course; another method is the information filtering method that gets the recommendation by filtering the students' information based on content.

The main technology used is the similarity analysis and then compare with the students' described file to find out the students' interest, this method is simple and effective but the shortcoming is not easy to find the new interest of students. So the idea of this paper is using this method to recommend the same or similar courses, it offers the evaluation of question and homework and examination results to find the knowledge that students don't master so as to give the best recommendation at the same time that is called the new found interest.The students' information is reasoned using Bayesian network after analyzing the students' information by the recommended rule and information filtering method, and then the platform provides them with adaptive learning material according to the students' situation.

c) The English resource library

We have annotated English resources in accordance with the teaching material from the lowest level of English first to the highest level of English six according to the students' characteristics. The platform provides the open generation tool of English grade reference library that the teachers can automatically generate the library reference, but the material is only provided for every student.

2) The personalized English automatic module

The module will automatically select the most matched personalized English learning according to the student's current English abilities from the English knowledge database. The basic principles include: providing students with the self-setting range adaptability material which is slightly more difficult than the student's language ability; providing the students with the English materials in forms and themes that are similar to the materials what the students are learning; providing the students with the current reading material to ensure it is the newest.

D. The results of the platform

We pick up 100 students who have different English abilities for using this platform for half a year, about 72% of the students' English abilities are greatly improved, about 13% of the students' English abilities are a bit raised .

V. THE CONCLUSION

The big data applying on the online English education platform of Guizhou realize the students' individual learning, it will not only achieve the equalization of resources to reflect the thought that take the students as origin, but also maximize the use of the resource of the advantage to enhance the quality of Guizhou students' English abilities. We will improve the level mark of English resources to enhance the platform's adaptability with the further development of the natural language technology.

REFERENCES

[1] Jiang Qiang, Zhao Wei, "Realization of Individual Adaptive Online Learning Analysis Model Based Big Data", Chinese educational technology, vol. 336, pp. 85–92, 2015.

[2] Ma Xiaoting, "Construction of the Big Data Analysis Platform for the Library based on the Personal Services Requirements", New practical library, vol. 6, pp. 20–23, 2014.

[3] GONG Xia-yi, "Survey on Big Data Platform Technology", journal of system simulation, vol. 26, pp. 489–496, 2014.

[4] Liu Zhi-hui, "Research overview of big data technology", journal of Zhejiang university(engineering science), vol. 48, pp. 1-14, 2014.

[5] Yanli Song, "The Construction of College Profession Network Resource based on Individuality" ,China adult education, vol. 22, pp. 42–43, 2009.

A method for water resources object identification and encoding based on EPC*

Ping Ai, Chuansheng Xiong, Hengli Liao, Dingbo Yuan, Zhaoxin Yue
Hohai University
Nanjing, China
e-mail: aip@hhu.edu.cn
e-mail: xcs123@163.com
e-mail: 48565752@qq.com
e-mail: 469143069@qq.com
e-mail: yzx10000@163.com

***Abstract*—The development and application of information technology largely expand the spatio-temporal scale and the element type of water resources information. With the increased amount and type, the water resources data is representing the characteristic of big data. According to the application requirements of the Water Resources Data Center, an encoding scheme of the Internet of Water Objects (WID) is designed based on the EPC (Electronic Product Code) is presented for processing the water resources big data.**

Keywords—Water resources data center; Water resources big data; EPC encoding; WID

I. INTRODUCTION

The development of social economy and technology has extended the water resources data service field, the application of the modern water resources data has not already limited to the traditional application areas, such as engineering design, disaster prevention and reduction, etc. On the while, the development and application of RS, GIS, GPS, IOT and other modern information collection technology has expanded the spatial and temporal scale and types of elements of the water resources data, making a sharp increase in quantity and types, showing the characteristics of multi-source, heterogeneous, massive. How to effectively store and apply the water resources big data has become one of key technical problems in water resources informatization[1-4].

How "big" are water resources data? Can be roughly estimated as follows: [5]

1. From the perspective of information collection: at the end of 2012, all over the country can receive provincial measurement (monitoring) stations about128,000, and generated data may be 60TB in real time in one year in China.

2. From the perspective of the first the national water resources census: involving 99 million water resources objects, and 400 million basic data records, which is a historical record. Assuming 5.5 million water resources objects and 560,000 social and economic water users need real-time monitoring, it will produce real time data about 2,781 TB in one year, according to each water resources object produces 16 bytes per second.

According to the water resources informatization planning, the construction of Water resources data center aims to fully integrate all kinds of scattered information resources, realize resources sharing, and carry on data mining deeply, in order to meet the needs of the development of water resources business (transaction.)[6]. However, the technical location and basic system of the existing Water resources data center lacks of data processing and application for massive water resources data (PB, even higher), especially semi-structured, and unstructured data, such as images, data flow, etc.

Aiming at the shortcomings of the traditional Water Resources Data Centre for processing water resources big data, this paper proposes an encoding scheme of water objects based on EPC to complete WID, to achieve information organization on objects and application on the condition of big data and finish the Water resources data center seamless integration with the IOT applications.

II. WATER RESOURCES OBJECTS AND THEIR IDENTIFICATIONS

A water resources object is a collection of bounded information, which describes a management object of water resources, whose basic characteristic is to use information to describe the properties and behavior of management entity of the water resources, and the properties to change only by their own behaviors. The state of the water resources object means a collection of all the attributes, which change with the change of attribute value. Identifying on the water resources object, essentially gives unique code to the object of water resources. The management object of water resources may be a physical entity (such as hydraulic engineering), can also be a logical entity or problem domain (such as implementation effects of the livelihood water policy).

The basic pattern of data organization of the Water resources data center is to organize data together on the same properties of the same objects, and this is the basic data model

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.73

of the relational database, according to the characteristics of the online transaction processing and applications, to achieve information classification and encoding. The identity information is information classification and encoding. In order to meet needs of information organization on objectification and application on the condition of big data, we need to achieve classification and encoding on the management objects of water resources, identify and organize data in the Water resources data center, for the function extension of the Water resources data center and meeting the need of the development of information technology.

According to the objects to organize information in the Water resources data center, it is necessary to organize all attributes information of objects from the given management objects of water resources. On the basis of the pattern of relationship data, the process of information organization on objectification is the process of information extraction and data storage through mapping relationship on the ID of information and objects.

With the development of water resources informatization, the application of the IOT technology will rapidly spread in the water conservancy industry, and the Water resources data center has to face the complex processing and application of sensor information from the internet of water objects. Therefore, using the identifier based on IOT to identify the water resources objects will form seamless integration the Water resources data center and the application of IOT, which will offer a good foundation for the management and applications of IOT based on WRDC.

At present, the globe has widely applied EPC for identification and encoding based on IOT. Research on the identification and encoding based on IOT in China, is mainly based on EPC system. Therefore, choosing the EPC code as the basis of identification and encoding for WID can be compatible with national and international standards in the future.

The characteristics of EPC lie in applying general encoding scheme to every item for encoding, this kind of encoding scheme involves just for identification of items, does not involve any features of the items, and each EPC code of the items in the IOT is equivalent to an index. EPC code consists of a set of numbers, including Header and other three pieces of data (General Manager Number, Object Class and Serial Number) [7]. EPC Header identifies the type and appoints the different length of the EPC code; General Manager Number describes the EPC information associated with the production of goods or management, such as "Guangdong Water Resources Data Center"; object classification records their exact type of information, such as "hydro-junction"; Serial Number uniquely identifies some given object under the domain administrator, For example ”the information stored in Feilaixia hydro-junction project of Guangdong Water resources data center”.

III. AN ENCODING SCHEME OF WATER OBJECTS BASED ON EPC TO COMPLETE THE WID

At present, the EPC code includes 64-bit, 96-bit, and other structures. In order to ensure that all water objects have their own EPC codes and be compatible with IOT outside the water resources industry, using 96-bit structure as the encoding scheme of water objects to complete WID. The 96-bit structure can offer 268 million unique identifications for General Manager Number, which can have 16 million types of object, and the type of each object can have 68 billion serial numbers, which is sufficient to meet the needs of construction and applications of the internet of water objects in China.

The general structure of the EPC encoding is a bit string, consists of a Header with hierarchical, variable length, and a series of digital field. The length, structure and function of the code are completely determined by the value of the Header. The Header defines the total length, identification types (function) and EPC encoding structure. In order to maintain maximum compatibility, the WID adopts general GID-96 as the basic encoding format. The General Identifier is defined for a 96-bit EPC, and is independent of any existing identity specification or convention. The General Identifier is composed of three fields -- the General Manager Number, Object Class and Serial Number. Encodings of the GID include a fourth field, the Header, which guarantees uniqueness in the EPC name space, as shown in Table 1[8].

Table 1. The General Identifier (GID-96)

	Header	General Manager Number	Object Class	Serial Number
GID-96	8bit	28 bit	24 bit	36 bit
	00110101（Binary value）	268,435,456（Decimal capacity）	16,777,216（Decimal capacity）	68,719,476,736（Decimal capacity）

According to the GID-96 and the features of water resources business applications and information processing and management, the basic encoding scheme of water objects based on EPC to complete the WID as follows:

1. GID-96 Header is fixed at 8bit, hexadecimal value is 35.

2. The General Manager Number is defined as the unique codes of the National WRDC nodes at all levels.

3. The Object Class is defined as the General Manager Number of the National Water resources data center nodes at all levels. The Object Class is defined by the nodes at all levels in accordance with their own needs, but to maintain uniqueness, and the default value is 0.

4. The Serial Number is defined as water resources objects. When define Object Class, the Serial Number should be unique within the category, when not define Object Class, the Serial Number should remain unique within the nodes.

The encoding structure of the WID is shown in Figure 1.

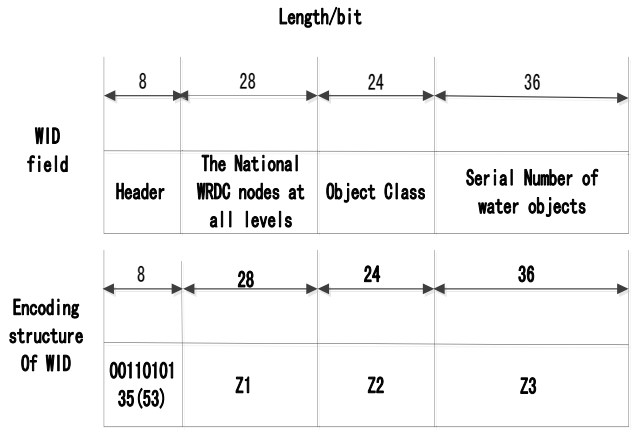


Fig.1 The coding structure of the ID of the internet of water objects

In figure 1 (each field is expressed in hexadecimal number for convenient), the WID code of each field is defined as follows:

1. The Header is a fixed code 35;

2. Nodes of the National Water resources data center at all levels are seven hexadecimal numbers (XXXXXXX);

3. The Object Class is six hexadecimal numbers, default is 000000, and nodes of the National Water resources data center at all levels may accomplish encoding according to their own needs;

4. The Serial Number of water objects from nodes of the National Water resources data center at all levels, are nine hexadecimal numbers (XXXXXXXXX);

Each WID code of water objects is represented by 24 hexadecimal numbers, when no Object Class code, the structure is 35XXXXXXX000000XXXXXXXXX.

For readability, the definition of equivalent decimal representation of WID is as follows:

53.Z1.Z2.Z3

Where:

53 is the decimal representation of the fixed Header;

Z1 is the decimal representation of nodes of the National Water resources data center at all levels, ranging from 0 to 268,435,456, and the maximum length is nine decimal data;

Z2 is the decimal representation of the Object Class, and the default value is 0, ranging from 0 to 16,777,216, and the maximum length is eight decimal data;

Z3 is the decimal representation of the Serial Number of water objects from nodes of the National Water resources data center at all levels, ranging from 0 to 68,719,476,736, and the maximum length is eleven decimal data;

Each field separated by "." cannot delete. When each field converts from the hex to decimal, leading zero is not recorded in the decimal representation string. Therefore, the decimal representation of WID is not more than 33 characters.

Since the government of China has not yet develop appropriate standards, combined with the code of the central government departments, this paper refers to the national industry code, and predefines reference codes (decimal) of the National Water resources data center at all levels shown in Table 2, when the introduction of national standards, which can replace the mentioned predefined nodes[9].

Table2. Reference code (decimal) of the National Water resources data center nodes at all levels

Ministry of Water Resources of the People's Republic of China 332	Changjiang Water Resources Commission of the Ministry of Water Resources 491	Yellow River Conservancy Commission of the Ministry of Water Resources 492
The Huaihe River Commission of the Ministry of Water Resources,P.R.C. 493	Haihe River Water Conservancy Commission,MWR 494	Pearl River Water Resources Commission of the Ministry of Water Resources 495
Songliao Water Resources Commission of the Ministry of Water Resources 496	Taihu Basin Authority of Ministry of Water Resources,P.R.CHINA 497	Beijing 1100
Jilin 2200	Jiangsu 3200	Fujian 3500
Henan 4100	Guangdong 4400	Sichuan 5100
Yunnan 5300	Gansu 6200	Xinjiang 6500
Tianjin 1200	Inner Mongolia 1500	Heilongjiang 2300
Zhejiang 3300	Jiangxi 3600	Hubei 4200
Guangxi 4500	Chongqing 5102	Tibet 5400
Qinghai 6300	Hebei 1300	Liaoning 2100
Shanghai 3100	Anhui 3400	Shandong 3700
Hunan 4300	Hainan 4600	Guizhou 5200
Shaanxi 6100	Ningxia 6400	Shanxi 1400

Since the definition of Object Class of water resources is much complex, according to the basic principles of the EPC, each current node cannot consider to be classified, or defines their Class to meet needs of their business. For compatibility with Object Class of the future national or water industry, Serial Number of water objects should remain unique within the scope of each node.

Encoding sample:

For example, encoding a hydrological station in Guangdong province: Header fixed for GID-96, is 00110101 (binary), and converts to hexadecimal is 35; The node of Water resources data center is Guangdong Water resources data center, the 10 hexadecimal code is 4400, and converts to hexadecimal is 1130; 000000 is no Object class; 00000001A is the serial number of WID for the hydrological station of the Guangdong Water resources data center (shown in figure 2).

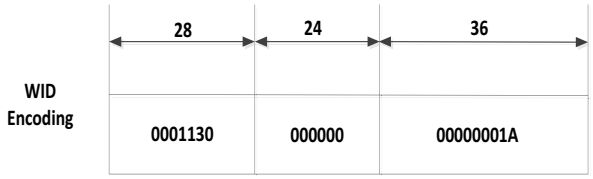


Fig.3 An encoding example of the hydrological station in Guangdong province

The WID hexadecimal encoding of the hydrological station in Guangdong province is represented as:

350001130000000000000001A

The corresponding decimal encoding is represented as: 53.4400.0.26.

IV. CONCLUSION

With the rapid development of information acquisition technology and network, especially aerospace remote sensing and IOT bring in "widespread perception of information", and make the water resources information both quantity and types be continuously enriched, and the water resources big data era is coming. This paper puts forward an encoding scheme of the Internet of Water Objects (WID) based on the EPC (Electronic Product Code) to organize information in the Water resources data center, and organize all attributes information of objects from the given management objects of water resources, which offers conditions for processing the water resources big data.

ACKNOWLEDGMENT

This research was supported by the Major Program of the National Social Science Foundation of China under Grant No. 2012&ZD214; the Major Research Plan Training Project of the National Natural Science Foundation of China under Grant No. 90924027; the Key Project of the National Natural Science Foundation of China under Grant No.41030636; the Graduate Research and Innovation Projects in Jiangsu Province under Grant No. CXZZ13_0261.

REFERENCES

[1] Kumar, S., Peters-Lidard, C., Tian, Y., Reichle, R., Geiger, J., Alonge, C., ... & Houser, P. (2008). An integrated hydrologic modeling and data assimilation framework. Computer, 41(12), 52-59.

[2] Abbott, M., & Vojinovic, Z. (2009). Applications of numerical modelling in hydroinformatics. Journal of Hydroinformatics, 11(3-4), 308-319.

[3] Gourbesville, P. (2009). Data and hydroinformatics: new possibilities and challenges. Journal of Hydroinformatics, 11(3-4), 330-343.

[4] Jonoski, A. (2012, September). Hydroinformatics and Decision Support: Current Technological Trends and Future Prospects. In BALWOIS 2012.

[5] Information center of Ministry of water resources. The 2012 annual Chese Water Resources Informatization Development Report[R].Beijing: Ministry of Water Resources Informatization Leading Group Office, 2013.

[6] Abbott, M. B. (1991). Hydroinformatics: information technology and the aquatic environment. Avebury Technical.

[7] Chen, X. Y., & Jin, Z. G. (2012). Research on key technology and applications for internet of things. Physics Procedia, 33, 561-566.

[8] EPCglobal. EPC Tag Data Standards Version 1.1 Rev.1.24. 2004,4.

[9] Industry code query. http://www.cye.com.cn/gongju/hangyedaima.htm.

Ensembling Base Classifiers To Improve Predictive Accuracy

Wen Qingdi
Department of Information
Guizhou University of Finance and Economics
Guiyang China
306873439@qq.com

***Abstract*---The algorithm of ensembling base classifiers can improve predictive accuracy, and achieve a better generalization. However, the ensemble classificition methods in literature have been used in more rule-based algorithms of classifier. This paper presents a novel algorithm: CVCEEP (Classification by Voting Classifiers based on Essential Emerging Patterns). By learning the method of Bagging, multiple base-classifiers were generated on different bootstrap samples and combined as a powerful classifier by voting. Experimental results show that CVCEEP achieve a better predictive accuracy and can be match to the classic classification algorithms that we have known.**

Keywords: ensemble learning, classification, emerging patter

I. INTRODUCTION

Classification is one of the most important technologies in data mining, many algorithms have been proposed so far. Classification is based on the characteristics of the data set to construct a classifier, then use of the classifier to the unknown class of samples to give the category. The process of constructing classifier is generally divided into two steps: training and testing. During the training phase, characteristics of training data set, generating an accurate description or model of the training data set. In the test phase, use of a description or model to classify the test date, test the classification accuracy.

There are many algorithms for solving classification problem in data mining. A good classification system should have a good generalization ability (a good classification ability for the new data, while avoiding the overfitting problem). Researches show that ensemble learning[1] can effectively improve the generalization ability of the classifier. The main idea of the algorithm is building the base classifier by analyzing or learning from a training set, then Ensembling these base classifiers to an efficient and accurate classification system.

This paper presents a Classification by Voting Classifiers based on Essentia Emerging Patterns, CVCEEP. The conventional classification algorithm for micro decision only consider one or a set of attributes of the way ,so we choose CEEP algorithm to Construct the base classifier. CEEP makes decision based on multiple attributes, so that it can improve the accuracy of base classifiers. By learning the method of Bagging, multiple base-classifiers were generated on different bootstrap samples, in the classification step, classifiers combined as a powerful classifier by voting. The Experiment study carried on benchmark datasets taken from the UCI machine learning repository show that CVCEEP achieves a better predictive accuracy and can be match to the classic classification algorithms such as C5.0, CBA, CMAR, CAEP and has excellent generalization ability.

II. ENSEMBLE CLASSIFICATION ALGORITHM

This section describes the main ideas and steps of the CVCEEP algorithm, The main content is the construction of the base classifier, training samples preprocessing, in the classification step, how to Ensemble the multiple classifiers.

A. Emerging Patterns

Assume the original data instances have m attribute values. Each instance in the training dataset D is associated with a class label, out of a total of n class labels:C_1, C_2, ..., C_n. We partition D into n sets, D_1, D_2, ..., D_n, with Di containing all instances of class C_i.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.74

Emerging patterns are defined for binary transaction databases. To find them, we may need to encode a raw dataset into a binary one: We discretize the value range of each continuous attribute into intervals. Each (attribute; interval) pair is called an item in the binary (transaction) database, which will be represented as an integer for convenience. An instance t in the raw dataset will then be mapped to a transaction of the binary database: t has the value 1 on exactly those items (A; v) where t's A-value is in the interval v. We will represent this new t as the set of items for which it takes 1, and we will assume henceforth the datasets D, D_1, ..., D_n are binary.

Definition 1 Let D be a subset of the training data set DB. Support for X on D is supD(X) = $count_D(X)/|D|$, countD(X) is the number of samples which contain X, | D | is the total number of samples in D. If D is a collection of C_i class training samples, supD(X) is labeled supi(X). That is the frequency of the X in the C_i class of training samples.

Definition 2 Let D and D' be a subset of the training data set DB. Set X from D 'to D grD' growth rate D (X) is defined as follows:

$$gr_{D\to D'}(X)=\begin{cases}0 & \text{if } sup_D(X)=sup_{D'}(X)=0\\ \infty & \text{if } sup_D(X)=0,\ sup_{D'}(X)\neq 0\\ sup_{D'}(X)/sup_D(X) & \text{else}\end{cases}$$

If the data set D and D' are Ci and non Ci' samples of the collection, grD'→D(X) is labeled gri(X), It is a measure of the degree of the change in the degree of support (frequency) of the X from the non Ci class to the Ci class.

Definition 3 Given the threshold of growth rate ρ >1, If the X is from D' to D growth rate grD'→D(X) ≥ ρ, X is called from D' to D EP, X is EP of D. Item set X is eEP of D, must meet the following conditions:

- X is EP's of D.
- The support level of X in D is not less than the minimum support threshold for the specified minimum support threshold ξ.
- Any sub set of X does not meet the conditions.

When D and D' are the sets of Ci and non Ci samples, EP/eEP's D is also known as the Ci class EP/eEP,, In fact, eEP is the most expressive EP.

B. The choice of the base classifier

In this paper, we use the CEEP algorithm to construct the base classifier. In the CEEP algorithm, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase). The training set is a set of items, the item is due of the attribute and the attribute value. Mining EP is to find the set of item which support in the two target classes is jumping. So the EP has very good classification ability. In the classification phase, for each target class C_i which have own EPs, then we aggregate the power of the discovered EPs for classifying an instance s: we derive an aggregate differentiating score for each class C_i, by summing the differentiating power of all EPs of C_i that occur in S; the score for C_i is then normalized by dividing it by some base score of the training set of C_i. Finally, we let the largest normalized score determine the winning class. The eEP is a special kind of EP. Its quantity is smaller than EP, but also has the strong ability of classification. The basic concepts of EP and classification algorithm based on EP is not the focus of this paper, we don't repeat them here, please refer to reference[2].

C. Create the base classifier.

According to the integrated algorithm, The parallel independent training set DB_1,DB_2,...,DB_λ are taken from the same training dataset D by bootsraping:

- The samples of DB_i are obtained from the dataset D by sampling with replacement.
- DB_i contains the N samples, where N is the total number of samples in training data set D.

There are some differences in the bootstrap samples sets DB_1,DB_2,...,DB_λ, because of using the random sampling. The samples are from the same data set D, so they reflect the same underlying data distribution.

The base classifiers are labeled as: BC_{B0}, BC_{B1}, BC_{B2},..., $BC_{B\lambda-1}$,the BC_{B0} is the classifier which directly learns from the original training set D. The Others are learned from the bootstrap sample set. From the training data set D, get λ-1 bootstrap sample sets DB_1,DB_2,…,$DB_{\lambda-1}$,then building the classifier labeled CB_m on the data sets DB_m ($1 \le m \le \lambda - 1$).The

λ classifiers(BC_{B0}, $BC_{B1}, BC_{B2}, ..., BC_{B\lambda-1}$) constitute a final classifier "committee", labeled CS(DB). For any unknown sample, each base classifier makes an independent prediction. The prediction results of each basis classifier can then be made to determine which target class the sample belongs to. The result is determined by BC_{B0} $BC_{B1}, BC_{B2}, ..., BC_{B\lambda-1}$, not one of them.

D. Integrated mode of multiple classifiers.

From the construction of the basis classifier in the upper section, $BC_{B1}, BC_{B2}, ..., BC_{B\lambda-1}$ is parallel. These classifiers are generated by different bootstrap samples, which are generated from the data set D according to the random way. There is no order relationship between the basis classifiers, so the prediction result of each classifier has the same weight 1. Although the classifier BC_{B0} is obtained from the original training data set, in the voting process it still has the same weigh. Only in the case of the base classifier is invalid, then the target class of the sample is determined by the BC_{B0} forecast. Ultimately, the results of the prediction of multiple base classifiers are combined by voting.

The training data set D contains $C_1, C_2, \ldots, C_K$ target class. The main classifier has λ basis classifiers. The integration of all the base classifiers is denoted as CS(DB). In order to classify the unknown sample S, we need to calculate the score of S belonging to the target class C_i, Vote(S,Ci). Let VBC(S)={BC|BC belongs CS(DB),and BC covers the sample S}. VBC(S) is a collection of voting qualifications base classifiers. Because the base classifier which failure of the sample S can't vote, so the VBC(S) is a collection of the real role base classifiers for the prediction of the unknown sample S. Vote(S,C_i) is the number of VBC (S) base classifiers can predict the sample of S, Vote(S,C_i) can be calculated by the voting process.

For any unknown sample *S*, the voting procedure is described as follows:

- For i=1,2,..,K, let Vote(*S,Ci*)=0.
- Calculate the base classifier set VBC(*S*).
- If VBC (*S*) is empty, then report the abnormal situation, the end of the voting process.
- For all the base classifiers BC_{Bm}€VBC(*S*), if BC_{Bm} predict the class *Ci* $(1 \le i \le K)$, then Vote(*S,Ci*)=Vote(*S,Ci*)+1.
- For i=1,2,..,k,Vote(*S,Ci*), record the number of votes obtained by the *Ci* class, return Vote(*S,Ci*). The voting process is end.

E. CVCEEP algorithm description

Classification by Voting Classifiers based on Essential Emerging Patterns, CVCEEP algorithm is described as follows:

Input: the date set *D*, the number of base classifiers λ.

- the minimum support threshold I and the minimum growth rate threshold ρ_I.
- Copy the training data set *D*, labeled as D_{B0}.
- For m=1…λ-1, construct self-sample set D_{Bm}.
- We build the λ base classifiers: $BC_{B0}, BC_{B1}, BC_{B2}, \ldots, BC_{B\lambda-1}$ from the training data sets: $D_{B0}, D_{B1}, D_{B2}, \ldots, D_{B\lambda-1}$,using the minimum support threshold I and the minimum growth rate threshold ρ_I. The BC_{Bm}($0 \le m \le \lambda - 1$) of the base classifier is trained by the corresponding training set D_{Bm}.
- For i = 1, …, k, in accordance with the voting process of the C section in the part 2, the votes Vote (S, Ci) is calculated.
- The *S* is classified to the most votes class. If the class is not only the highest number of votes, .the votes of the highest class *S* under the most. If all base classifiers fail, the *S* under the BC_{B0} judgment as the final prediction.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we report CVCEEP algorithm and compare with the C5.0,CBA, CMAR, CAEP, and CEEP as the basis of, and analyze the classification of CVCEEP algorithm.

A. Comparison experiment of CVCEEP algorithm

Table 1 gives the correct rate of various classification methods in the 12 datasets. At the table, "_" means the lack of data, the average value is the average of the classification accuracy of all algorithms. Experimental results show that the CVCEEP has 10

wins in 12 data sets and has the best average accuracy. In the Pima data set, the enhancement effect is the most obvious. The classification accuracy is 5.5 percentage points compared with the algorithm based on CEEP. This shows that the method of constructing multiple classifiers and voting based classifier is reasonable and effective.

B. CVCEEP algorithm performance analysis

The number of the base classifiers λ is an important parameter of CVCEEP algorithm. After repeated experiments, the accuracy of CVCEEP algorithm can be significantly improved with the increase of the number of basic classifiers. After a certain number of classifiers, the accuracy rate will be stable in the higher region. Experiments show that usually in the number of base classifiers in between 15 to 35 CVCEEP algorithm classification accuracy rate will reach the stable region. After the number of the base classifier reached 41, the accuracy is relatively stable.

TABLE I. CLASSIFICATION ACCURACY: C5.0, CBA, CMAR, CAEP, CEEP AND CVCEEP COMPARISON

Data Set	C5.0	CBA	CMAR	CAEP	CEEP	VCEEP
Adult	85.54			83.09	81.85	82.04
Australian	84.93	84.9	86.1	86.21	87.83	89.42
Cleve	77.16	82.8	82.2	83.25	84.33	88.67
Diabete	73.03	74.5	75.6	67.3	74.08	77.24
German	71.9	73.4	74.9	72.5	73.4	76.4
Heart	76.3	81.9	82.2	83.7	84.07	86.67
Iono	91.45	92.3	91.5	89.76	92.57	93.14
Mushroom	100			98.82	99.9	100
Pima	75.39	72.9	75.1	75	73.03	78.55
Sonar	76	77.5	79.4	78.3	85.5	86
Vechile	69.82	68.7	68.8	66.32	67.44	71.79
Lymph	78.29	77.8	83.1	74.38	84.38	85
average	79.98	80.86	81.88	79.89	82.37	84.51

Ⅳ. CONCLUSIONS

This paper presents a classification algorithm based on eEP for multi classifier voting, CVCEEP. It combines the advantages of EP based classification and ensemble method. The algorithm has achieved good results in the UCI machine learning library. The classification accuracy of CVCEEP algorithm is comparable with the existing classification algorithms. The parallel design of the algorithm is easy to implement, and has strong scalability.

REFERENCES

[1]. Dietterich T,Thomas G.Machine learning research:Four current directions.AI Magazine,1997,18(4):97?136.

[2]. G. Dong, X. Zhang, L. Wong, and J. Li. CAEP: Classification by Aggregating emerging patterns. In Proc. of the 2nd Int'l Conf. On Discovery Science (DS'99), pp 30-42, Tokyo, Japan, Dec. 1999.

[3]. Valiant L. G.A The theory of Learnable. Communication of ACM'27 Pages 1134-1142, 1984

Expert Achievements Model for Scientific and Technological Based on Association Mining

Xuexin Qu, Rongjing Hu
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, Jiangsu Province, China
245050598@qq.com

Lei Zhou, Liuyang Wang and Quanyin Zhu *
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, Jiangsu Province, China
hyitzqy@126.com

***Abstract*—In order to improve the use value of expert information, the Information technology experts Mining System is designed and realized which based on Mining Association. University website on the Internet, CNKI and Soopat patent website to provide expert information related data. Through the application system designed based on Web crawling, association mining model and page analysis, good results achieved by digging the expert information from universities and first-class hospitals in Jiangsu. According to the data mining from the database of SooPAT and CNKI, correlation algorithm model of the technology experts mining system is designed and achieved 92% of the associated information accuracy and 97% of the expert information accuracy in experiment, which provided the data support for the system design. Available to the government departments or technology enterprises to browse, the system can meet the demands of the information of scientific and technological information of technology experts of government science and technology departments and science and technology enterprises.**

Keywords- Expert information; Association mining; Web mining; Web crawler

I. INTRODUCTION

With the development of Internet and technology, Scientific and technological enterprises expand the demands for experts. At home and abroad, Web mining method already developed a lot of methods to obtain expert information. Most of the expert information mining is just a single access to personal information and results of information, including journals, Conferences, patents information, and so on. This way only provide access to single paper information or expert information, which cannot do associated analyze and unable to meet science and technology enterprises and government on the collection of scientific and technological achievements. This paper proposes an improved expert information mining patterns, which increases the degree of perfection of expert information, meets the demand for Government sector's and high-tech enterprise's needs of Research results data collection and consultancy services. Data used in this paper is expert information in university and the first-class hospitals in Jiangsu province, which also associated the information of papers and patents of the experts. Experimental results show that, the more the expert information's result improved, the more the result of the associated information accurate.

*Corresponding author. hyitzqy@126. Com

II. ASSOCIATION MINING MODEL

Establish data set $Url=\{u_1,u_2,\ldots,u_m\}$,we use it to storage link, Establish expert dataset $sc=\{s_1,s_2,...,s_n\}$, which has n elements to storage information of experts, Establish a set of expert papers $paper=\{a_1,a_2,...,a_n\}$, which has n elements to storage information of papers, Establish expert patent dataset ,$patent=\{p_1,p_2,...,p_n\}$, which has n elements to storage information of patent. The model is shown in Figure 1.

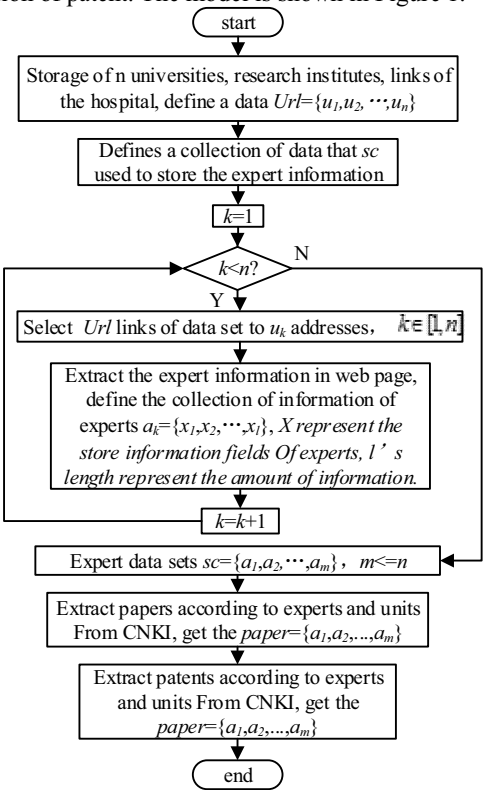


Figure 1. Association mining model

Extraction the information of experts by *Url* set *sc* that is consist of name, positional titles, department, unit. Then, the information set of $paper=\{a_1,a_2,...,a_n\}$ is gotten according to the name and place from http://www.cnki.net/, Similarly, extraction the information set of patent from http://www.soopat.com/ according to experts' name and place.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.75

III. EXPERIMENTAL ENVIRONMENT

In the test article, written using the Python language Web crawler, which, BeautifulSoup to analyze web pages to fetch data, using Mysql database, and ultimately dig papers from known online. The following table shows the configuration table for the operation of the computer where the reptiles.

TABLE I. EXPERIMENTAL ENVIRONMENT

CPU	Memory	OS	Python	Mysql
Intel I5	4G	Windos7	Python3.4	Mysql 5.6.10

Accuracy and experts associated with the accuracy of the experiments in this paper based on the mining association, the experts test data. Raw data, the associated data acquired from CNKI and Soopat patent website to provide expert information and universities in Jiangsu Province from three hospitals. Extracting 10 units were testing samples expert information, the accuracy of testing experts and related information accuracy.

IV. EXPERIMENT WITH DIFFERENT DATA

A. *Web crawler test*

The crawler test program from 18:10 began to run, three crawlers running at the same time, a public IP in CNKI climb take expert thesis data, eventually in one hour twenty minutes to get to 80000 more data. Please consult Table 2.

TABLE II. THE BEST AVERAGE ERROR OF THE TEN KINDS OF AGRICULTURAL PRODUCTS

Time	current data size	mining data size
18:40	0	0
18:42	2061	2061
18:44	4208	2147
18:46	6495	2287
18:48	8923	2428
18:50	11127	2204
18:52	13500	2373
18:54	15983	2483
18:56	17524	1541
18:58	18848	1324
19:00	20235	1387

Reptile specific data as shown in Table 2, the following table can be seen with the increasing amount of data, the amount of crawled, slowly begins to decrease, we can see that the reptile on certain performance needs to be improved

B. *Expert information accuracy*

Prior to the test data, in order to better describe the results, some definitions are simplified. The schools and hospitals are defined be SC, units please consults Table 3.

TABLE III. THE LIST OF COLLEGES AND UNIVERSITIES

Test name	Unit name
SC1	Nanjing University of Science and Technology
SC2	Nanjing University of Aeronautics and Astronautics
SC3	Jiangnan University
SC4	Hohai University
SC5	Nanjing University
SC6	Jiangsu Provincial People's Hospital
SC7	Nanjing children's Hospital
SC8	Hospital of integrated traditional Chinese and Western medicine in Jiangsu Province
SC9	Jiangsu provincial hospital
SC10	Maternal and Child Health Hospital of Jiangsu Province

TABLE IV. 10 KINDS OF UNIT ACCURACY ANALYSIS

	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10
SC1	0.96	0.94	0.94	0.98	0.94	0.96	1	0.96	0.94	0.96
SC2	0.96	0.98	0.92	0.94	1	0.94	0.96	0.92	0.96	0.98
SC3	0.94	0.98	0.94	0.96	0.96	0.92	1	0.94	0.96	0.94
SC4	0.88	0.9	0.88	0.88	0.94	0.9	0.92	0.86	0.9	0.88
SC5	0.88	0.86	0.92	0.86	0.94	0.9	0.92	0.88	0.9	0.88
SC6	0.98	0.98	0.94	1	0.92	0.98	0.94	0.96	0.94	1
SC7	0.96	0.98	1	0.96	0.98	1	0.94	1	0.96	0.98
SC8	0.98	0.98	0.98	0.94	0.96	0.96	1	0.94	0.96	0.96
SC9	0.96	1	0.98	1	0.94	0.96	0.98	0.96	1	0.98
SC10	0.97	0.98	1	0.98	0.96	0.96	0.96	1	0.98	0.96

In table 4, the data is used in this experiment, the data gets through randomly.

In order to visually describe the data in table 4, Figure 1, shows the data in table 2 trends. By Figure 1 it can be seen that experts from Hohai University and Nanjing University information crawling rate did not reach more than 90%. Hospital accuracy is much better than schools, which crawl in Nanjing maternity and child health care hospital data is the most accurate, with an average of 97.5%, number of error is 5 people or less. Analysis of accuracy of school status at the end of, mainly, in the judging error generating, followed by school into secondary school, need more grab script. In general, the accuracy of expert information has reached the standard.

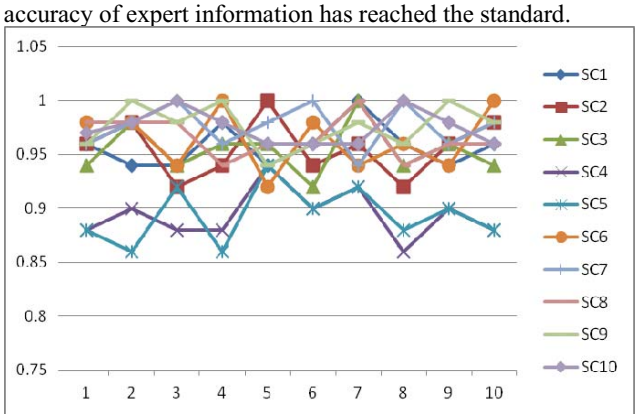


Figure 2. 10 kinds of unit accuracy analysis

In order to visually describe the data in table 4, Figure 2, shows the data in table 2 trends. By Figure 2 it can be seen that experts from Hohai University and Nanjing University

information crawling rate did not reach 90%. Hospital accuracy is much better than schools, which crawl in Maternal and Child Health Hospital of Jiangsu Province data is the most accurate, with an average of 97.5%, number of error is 5 people or less. Analysis of accuracy of school status at the end of, mainly, in the judging error generating, followed by school into secondary school, need more grab script. In general, the accuracy of expert information has reached the standard.

TABLE V. THE AVERAGE ACCURACY OF EXPERT INFORMATION

Min	Max	Avg	Avg1	Avg2	Avg3	Avg4	Avg5	Avg6	Avg7	Avg8	Avg9	Avg10
0.894	0.976	0.951	0.958	0.956	0.954	0.894	0.894	0.964	0.976	0.966	0.976	0.975

Min: The minimum value of the Min; Max: The maximum value of the Max; Avg: The overall average of these data; Avg(1-10): The average of the training sample data;

As is shown in Table 5, the average of test data in the table can be seen that the average accuracy rate of the expert information is 95.1%.

TABLE VI. RANDOM SAMPLE OF EXPERT INFORMATION

Test group 1	Test group 2	Test group 3	Test group 4	Test group 5
0.88	0.85	0.98	0.93	0.92
0.99	0.87	0.96	0.91	0.99
0.96	0.98	0.84	0.85	0.92
0.91	0.91	0.9	0.97	0.88
0.83	0.95	0.92	0.93	0.98
0.94	0.92	0.89	0.99	0.9
0.96	0.82	0.97	0.87	0.96

In Table 6, the expert information is randomly selected sample drawn five groups, shown in Figure 3. The accuracy rate for the expert information is shown in Figure 3. From this picture, we can see that the accuracy rate of expert information is basically the same as that in Figure 1. As is shown in Table 7, the accuracy rate of the random sample is 92%.

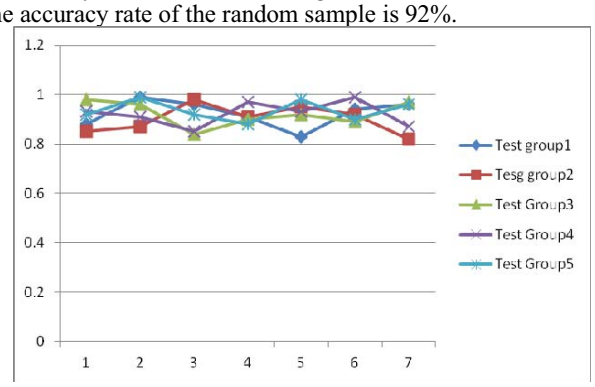


Figure 3. The best price forecast of beef under the three methods

TABLE VII. 10 KINDS OF AGRICULTURAL PRODUCTS DATA ANALYSIS

Min	Max	Avg	Avg1	Avg2	Avg3	Avg4	Avg5
0.9	0.9357	0.9208	0.9242	0.9	0.9228	0.9214	0.9357

Comprehensive two sampling patterns, according to table 6 and table 7, it can be drawn that the overall accuracy of the experts in 93%.

C. *Expert information associated with accuracy*

As is shown in Figure 3, the accuracy rate of a sample of expert information is random. Can be seen from table 8, of Association are accurate and reliable, apart from a few isolated associated data, accuracy is above 90%.

TABLE VIII. RANDOMLY ASSOCIATED EXPERT INFORMATION ACCURACY

A	B	C	D
0.9444	0.909	1	0.8
0.9756	1	1	1
0.5714	0.9545	1	0.9090
1	1	1	0.6875
1	1	1	1
0.9565	0.9886	1	1
1	1	0.9855	1
0.9841	1	1	1
1	1	1	1
1	1	1	1
1	0.8	1	1
1	1	1	0.6071
1	1	1	1
1	1	1	0.9047
1	0.8889	0.8333	0.9795
0.9807	0.9945	0.9677	1
0.625	0.9615	1	1
0.9642	1	1	1

A-D: Test procedure;

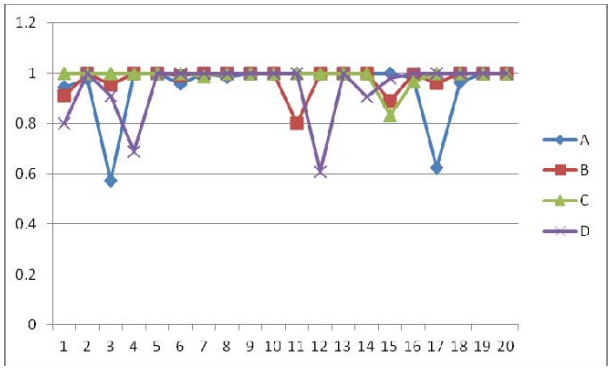


Figure 4. Accuracy analysis of random correlation expert information

As is shown in Figure 4, for expert information associated with accuracy chart, you can see in the figure, low accuracy rate in the number of 6, the other expert information associated with a more comprehensive.

TABLE IX. AVERAGE ACCURACY RATE OF EXPERT ASSOCIATION INFORMATION

Min	Max	Avg	Avg1	Avg2	Avg3	Avg4
0.9647	0.9859	0.9731	0.9663	0.9755	0.9859	0.9647

As is shown in table 9, for an average of associated experts of information accuracy, can be seen in the table accuracy is 97% associated with expert information.

Comprehensive shown in Figure 3 and Table 9， expert associate high accuracy results full of experts to meet and function of the Government sector and high-tech enterprises of science and technology expert research data collection.

CONCLUSION

The research data of retreatment comes from related work. According to the experimental results, we can see that the accuracy rate of expert information is 92%, and the accuracy rate of related information is 97%, which achieve good effect. This method needs further exploration and experimental for verification in other environment, such as increased digging units and experts of the same name in the same department processing method. In association process, the same department and namesake expert's information cannot be distinguished between treatments, so how to distinguish and process the data in the same department experts with the same name is another research direction.

ACKNOWLEDGMENT

This work is supported by the National Sparking Plan Project of China (2011GA690190), the fund of Huaian Industry Science and Technology. China (HAG2014023, HAG2014028).

REFERENCE

[1] Yi-Dong Shen, Zhong Zhang, Qiang Yang. Objective-oriented utility-based association mining. Proceedings of IEEE International Conference on 2002, 2002, 426 - 433

[2] Cinicioglu, Ertek, Demirer, Yoruk. A framework for automated association mining over multiple databases. Proceedings of 2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2011, 582 - 586

[3] Shoaib, Maurya. URL ordering based performance evaluation of Web crawler. 2014 International Conference on Advances in Engineering and Technology Research (ICAETR), 2014, 1 - 7

[4] Kan Hu, Cheung, David W, Shaowei Xia . Adaptive interval configuration to enhance dynamic approach for mining association rules. Tsinghua Science and Technology, 1999, Vol.4(1):1325 - 1333

[5] Jiawei Han, Fu. Mining multiple-level association rules in large databases. IEEE Transactions on Knowledge and Data Engineering. 1999, Vol.11(5):798 - 805

[6] Dahiwale, Raghuwanshi, Malik, L. Design of improved focused web crawler by analyzing semantic nature of URL and anchor text. 2014 9th International Conference on Industrial and Information Systems (ICIIS), 2014, 1 - 6

[7] Bing Liu, Grossman, Zhai. Mining Web pages for data records. IEEE Intelligent Systems,2004, 49 - 55

[8] Yuekui Yang, Yajun Du, Yufeng Hai, Zhaoqiong Gao. A Topic-Specific Web Crawler with Web Page Hierarchy Based on HTML Dom-Tree.2010 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA), 2010, 119 - 123

[9] Chan, Wei Fan, Prodromidis, Stolfo. Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and their Applications, 1999, 67 – 74

[10] Quanyin Zhu, Pei Zhou. The System Architecture for the Basic Information of Science and Technology Experts Based on Distributed Storage and Web Mining. Proceedings of International Conference on Computer Science & Service System. 2012, 661 - 664

[11] Quanyin Zhu, Hong Zhou, Yunyang Yan, Jin Qian and Pei Zhou. Commodities Price Dynamic Trend Analysis Based on Web Mining. Proceedings of Third International Conference on Multimedia Information Networking and Security. 2012, 524 - 527

[12] Quanyin Zhu, Jin Ding, Yonghua Yin, Pei Zhou. A Hybrid Approach for New Products Discovery of Cell Phone Based on Web Mining. Journal of Information and Computational Science. 2012, Vol. 9(16):5039 - 5046

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Novel Solution of Event Conflict Resolution Based on D-S Evidence Theory

Xiaojuan Yang
School of Communication
Shandong Normal University
Jinan, China
yxjuan08@126.com

Tao Sun* (*corresponding author*)
School of Informtion
Qilu University of Technology
Jinan, China
suntao0906@163.com

***Abstract*—Usually, there will still be data conflicts left in the event mentions after the events have been co-reference resolution. This paper presents a novel experiments-proven solution to the conflicts of event mentions by first categorize the conflicts of events into named entity conflicts, semantic conflicts and data presentation conflicts and then take advantage of the D-S evidence theory in dealing with uncertain data based on "evidence" and "combination" .**

Keywords- Event Coreference Resolution ; Data Conflict Resolution; Event Correlation; Market Intelligence Analysis

I. INTRODUCTION

The analysis of market intelligence rely on accurate, comprehensive and creditable data. However, event mentions that extracts from different data source often has different description for the same underlying real-world event. Some description can reflects all or partial information correctly, but some can be discrepancy and even in contradiction with the actual facts. Moreover, description themselves can be inconsistent with each other.

This paper firstly categorize conflicts of events into named conflicts, semantic conflicts and data presentation conflicts according to the features of events description and then propose a novel solution for conflicts of events and data based on D-S evidence theory to improve the accuracy for discovering the true value of events attributes. Experiments on multiple real data sets shows that our approach can fulfill the task of solving the data conflicts between web entity co-reference events and have relatively better adaptability and accuracy.

II. THE SOLUTION TO THE CONFLICTS OF EVENTS

A. Categories of conflicts of events

According to the features of events description, this paper categorize conflicts of events into three classes:

(1) Named conflicts. This kind of conflicts of events is mainly about conflicts of data, often exist in the participant's attribute facts including the subject and object of the event.

(2) Data presentation conflicts. This kind of conflicts of events is also mainly related to the conflicts of data, often exist in the time, location attributes facts.

(3) Semantic conflicts. This kind of conflicts of events exist mainly in the activity (motion) attributes.

Traditionally, we deal with the above three kinds of conflicts of events using preset conflict-handling rules. We always solve the conflict when it occurs. However, this kind of conflict-handing approach cannot adapt to the complex and changeful web environment so that preset rules are not always effective.

In the essence, the solution of event conflict resolution is an uncertainty reasoning issue. D-S evidence theory is a math method aims at dealing with uncertainty problems based on "evidence" and "combination". In this paper, in order to tackle with the complexity, connectivity and inevitability, we propose a novel solution to the conflicts of events based on D-S evidence theory.

B. Two-Stage Solution to Conflicts of Events

Base on the above analysis, we propose a two-stage solution to the conflicts of events attributes.

(1) Firstly, we calculate the conflicts of event attributes and categorize them. Specifically, we assign the conflicts arise from participants, time and location, type of events into naming conflicts, data presentation conflicts and semantic conflicts respectively.

(2) For each conflicts class, we calculate its conflicts extent. For those conflicts extent smaller than threshold, we use strategies such as voting mechanism or the compare of confidence value to get the partial true value of the events appearance in the first stage.

(3) The extent of conflicts of the remaining event attributes will be relatively larger since the first stage has exported attributes facts. In order to put the result of the first-stage conflicts handling into better use and improve the accuracy of the solution to the conflicts of event attributes data, we add the dependent relationship between data source and the relationship between data source and the actual facts and we aim to get the true value of the events attribute attributes in the second stage using the results from integration of different strategies based on D-S evidence theory.

(4) Finally, we combine the conflict-free event attribute facts resulting from the two-stage solution to the conflicts of event attributes into full and real events description record.

The input of the solution to the conflicts of event attributes come from different events set from different data source, and has been processed with web entity events co-reference resolution. The export is combination of events whose conflicts of event attributes have been solved by two-

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.76

stage approach. The algorithm of the above two-stage solution to the conflicts of event attributes is as follow:

ALGORITHM1: TWO STAGE SOLUTION TO THE CONFLICTS OF EVENT ATTRIBUTES

Input: Web entity co-reference event set E_C , including event attributes set Ar , event set E , threshold of data conflicts T ;
Export: event set E_R , in which the data conflicts in its attributes has been solved.

1. initialization: $E_R=\varnothing$;
2. According to the type of conflicts of events, categorize the event attribute into three class // first stage
3. For each type of conflicts of events, calculate the extent of event attributes.
4. Use voting and maximum confidence value of facts to solve the conflicts respectively and get the event set $E_R = E_R \cup \{e_k\}$.
5. for $e_i \in E$ // discover the true value in second stage
6. if $Conflict(a_i) < T$
7. $Ar_L = Ar_L \cup \{a_i\}$
8. else
9. $Ar_H = Ar_H \cup \{a_i\}$
10. solve the conflicts of event attributes in Ar_L and get result E_L .
11. E_H solve the conflicts of event attributes in Ar_H . Introduce new rules and get result E_H based on D-S evidence theory.
12. for $e_i \in E$
13. for $a_j \in A$
14. choose corresponding true value based on result set E_L 、 E_H
15. form record e_j for the corresponding true value for the attributes, E_R=E$_R \cup \{e_j\}$.
16. Return data set E_R .

To improve the efficiency of conflict solving, we categorize the attributes of the co-reference event mentions based on the type of conflicts of events. For those attributes fact that has smaller extent of conflicts, voting or statistical method can have relatively high accuracy. For those attributes fact that has larger extent of conflicts, we employ the idea of uncertainty reasoning and blending from D-S evidence theory to synthesize the confidence value of various vent attributes fact to get consistent result and information complementarity.

Since event mentions that come from different data source have significant difference on many aspects such as description style and logicality, thus the conflicts of event attributes can be very complicated. Usually, the more event attributes fact from different data source, the more the uncertainty is and the more the extent of conflicts is .To measure the extent of the conflicts of event attribute, we use information entropy to explain the extent of event attributes.

For the attribute of a given event $ea_i \in EA$ and the event set from different data set $F_{ea_i} = \{f_1, f_2, \cdots, f_M\}$, then the extent of the conflicts of event attributes ea_i can be defined as:

$$Conflict(ea_i) = -\frac{1}{\log M}\sum_{j=1}^{M} p(f_j) \bullet \log p(f_j) \qquad (1)$$

In which, $p(f_j)$ is the ex-ante probability of event attributes fact f_j , $p(f_j) = \frac{|f_j|}{\sum_{j=1}^{L}|f_j|}$; $|f_j|$ is for the event fact occurrence $f_j (1 \le j \le M)$, $\log M$ is normalized factor.

If the extent of the conflicts of event attributes from different data source has exceed the threshold, then the simple event attribute conflicts strategy will not gain enough accuracy so we must consider other factors such as the credibility of data source, the accuracy of the event attributes facts and the interdependence between data sets.

III. EXPERIMENT AND EVALUATION

We build three data set: enterprise events data set (ENDS), personal events data set (PEDS) and product events data set (PRDS). Each data set is composed of about 100,000 data. To ensure the quality of main data, the ratio of positive examples and negative examples has been designed as 1:4.

To test the effectiveness of the approach we proposed, we will analyze and evaluate the event conflict resolution based on D-S evidence theory from two aspects respectively: (1) the accuracy of different event conflict resolution strategy (2) the impact of different combinations of rules on the accuracy of event conflict resolution.

A. The accuracy of different event conflict resolution strategy.

The accuracy of the event conflict resolution strategy refers to the percentage of the event attributes facts that have correctly recognize the true value of event attributes on the total number of event attributes facts. We compare the accuracy of event conflict resolution based on D-S evidence theory (D-S ECR), TruthFinder[1]and the method mentioned in literature[2](Voting) using the three data set we built above. Note that since TurthFinder and Voting are all used in the entity data conflict, we revise the algorithm of the two method to adapt to our data sets and we exclude the event activity (verb) attributes when compared with D-S ECR.

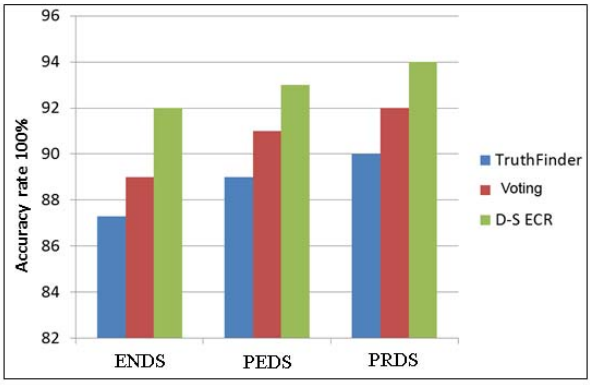


Figure 1. Accuracy rate of event conflict resolution

Figure 1 presents the comparison of our approach with the other two methods. We can see from Figure 1that our approach has relatively high accuracy (93.92%), D-S ECR has taken full advantage of various kinds of features such as text, data source, frequency of event occurrence and semantics and it combines multiple event conflict resolution strategy, which has provided the D-S ECR higher accuracy compared with that merely consider one factor such as credibility of data and data source and data dependency. Simple Voting is effective on majority of circumstances compared with TruthFinder in the experiment on three data sets. In conclusion, the D-S ECR approach can boost the accuracy of event conflict resolution effectively since it employs many conflict resolution strategy and multi-angle features and rules.

B. *The impact of different combinations of rules on the accuracy of event conflict resolution.*

To verify he effectiveness of the event conflict resolution approach we proposed in this paper, we test the impact of different strategy and combination on the accuracy. In the experiment, we test the voting strategy (V), interaction between data source and event attributes facts strategy (SF) and semantic factors (SD). We did not compare the location of test strategy because it can only be used in the first stage. Voting strategy is our benchmark strategy in the sense that the accuracy result of the other two strategies will be compared with the voting strategy to test the accuracy of event conflict resolution when combination of them are employed.

Figure 2 presents the accuracy corresponding to different strategy and the combination of the strategies. We can see that the strategy we proposed have all improved the accuracy of event conflict resolution to some extent using the voting strategy as benchmark, thus we have proved the effectiveness of the strategy we proposed. Moreover, the interaction between data source and event attributes facts strategy has also increased the accuracy remarkably, which confirms the existence of high quality (credible) data source and the improvement of accuracy from dependency of data source is limited due to the fact that there is no copy of large-scale data in the resolution phase in the event conflict and only small portion of small copy situation exists. And, we did not consider the direction of data dependency which is also a reason that the effect of data source dependency is not obvious.

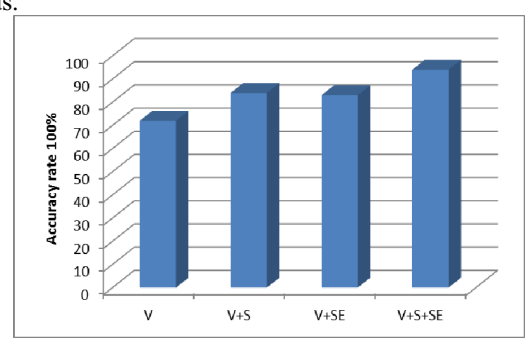


Figure 2. Impact of different event conflict resolution strategies on accuracy

IV. RELATED WORKS

Data conflict was first been raised in the database field[3] and the research was mainly focused on relationship extension where Naumann[4] and other have done fruitful work at adopting different conflict resolution strategy and function for different type of data conflicts. Several literature[5][6][7] have done review on the data conflict strategy in data integration area.

Besides the database field, data conflict resolution in Web area has mainly concentrated on the entity resolution of Deep Web, where we choose the true value from many conflict value to describe actual entity information correctly and comprehensively. Literature[2] believe that web data source has strong relevance and then analyze the interdependency relationship between data source to judge its credibility and to find the true value of different data source. Dong Xin[8] define and describe the dependency issue in detail. Literatures[9][10][11] further consider the accuracy of data source and combine it with dependency of data source to obtain fairly satisfying results. Li, Zhang [12][13]and etc. proposed a two state data conflict resolution based on Markov logic network which used multi-angle features and rules to improve the accuracy of data conflict resolution.

There are few study focused on the data conflict in the event area. Actually, we have not seen a study aimed to resolve the conflict using the features of event, entity, text and data source. We expect more attention will be paid on this issue since the acquisition of information is very essential to the analysis of market intelligence.

V. CONCLUSION

This paper proposed a two-stage data conflict resolution based on D-S evidence theory. In the first stage, we classify the event attributes based on the type of event conflict where we can obtain high accuracy just by using simple voting strategy for relative small extent event conflict. In the second stage, we using the credibility of data and data source, the dependency of data source and other featured to form data conflict resolution where we employ the extended combined rules of D-S evidence theory to integrate many conflict resolution strategy to improve the accuracy of event conflict resolution and allow the addition of new strategy and features. Experiments have proved that our approach has strong adaptation.

ACKNOWLEDGMENT

This work is supported by the projects of Shandong Province Higher Educational Science and Technology Program (J12LN20), China.

Tao Sun is corresponding author (*E-mail: suntao0906@163.com*)

REFERENCES

[1] X. L. Dong, L. Berti-Equille, D. Srivastava. Integrating Conflicting Data: The Role of Source Dependence [J]. Proceedings of the VLDB Endowment (PVLDB), 2009, 2(1):550-561.

[2] X. Yin, J. Han, P. S Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web [C]. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, 2007, pp.1048-1052.

[3] Dayal U. Processing queries over generalization hierarchies in a multidatabase system //Proceedings of the VLDB.Florence,Italy,1983:342-353..

[4] J. Bleiholder, F. Naumann. Conflict Handling Strategies in an Integrated Information System [C]. In Proceedings of the International Workshop on Information Integration on the Web (IIWeb). Edinburgh, UK, 2006.

[5] Papakonstantionou Y,Abiteboul S,Garica-Molina H.Object fusion in mediator systems//Proceedings of the VLDB.Bombay,India,1996:413-424.

[6] Motro A,Anokhin P,Acar A C.Utility-based resolution of data inconsistencies //Proceedings of the IQIS Workshop. Paris,France, 2004：35-43.

[7] Schallehn E,Sattler K-U,Saake G.Efficient similarity-based operations for data integration. Data and Knowledge Engineering, 2004，48（3）：361-387.

[8] L. Berti-Equille, A. Das Sarma, X. L. Dong, A. Marian, D. Srivastava. Sailing the Information Ocean with Awareness of Currents: Discovery and Application of Source Dependence [C]. In Proceedings of 4th Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, USA, 2009.

[9] X. L. Dong, L. Berti-Equille, Y. Hu, D. Srivastava. Global Detection of Complex Copying Relationships between Sources [J]. Proceedings of the VLDB Endowment (PVLDB), 2010, 3(1):1358-1369.

[10] X. L. Dong, L. Berti-Equille. Large-Scale Copy Detection [C]. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Athens, Greece, 2011, pp.1205-1208.

[11] X. Liu, X. L. Dong, B. C. Ooi, D. Srivastava. Online Data Fusion [J]. Proceedings of the VLDB Endowment (PVLDB), 4(11):932-943, 2010.

[12] Q L Li，YX Zhang, L Z Cui Data Conflict Resolution with Markov Logic Networks，m The SPRING 9th International Conference on Computing, Communications and Control Technologies, 2011.

[13] Y F Zhang, T Huang Markov Logic Networks with its application in De-duplication, m Journal of ChongQing University, 33，2010.

A Scene Analysis Model for Water Resources Big Data*

Ping Ai, Zhaoxin Yue, Dingbo Yuan, Hengli Liao, Chuansheng Xiong
Hohai University
Nanjing, China
e-mail: aip@hhu.edu.cn
e-mail: yzx10000@163.com
e-mail: 469143069@qq.com
e-mail: 48565752@qq.com
e-mail: xcs123@163.com

***Abstract*—According to the application requirements of the Water Resources Data Centre (WRDC), a scene analysis model is proposed for organizing and applying the objective information under the condition of big data. This model focuses on the generation and application of the digital scene, supporting big data processing, and establishing a new application mode of WRDC. This new model has the ability to combine the causal analysis and the association deduction. The advantage of the proposed method lies in its ability for rapid processing and dynamical analyzing the water resources big data, to solve problems in water situation evaluation and water resource decisions.**

Keywords—WRDC; Water resources big data; Information organization; Scene analysis

I. INTRODUCTION

The development of social economy and technology has extended the water resources data service field, and the development and application of RS, GIS, GPS, IOT and other modern information collection technology has expanded the spatial and temporal scale and types of elements of the water resources data, making a sharp increase in quantity and types [1-3].

Big data thought as a novel data processing technology, referring to the large-scale, complex structure and high intensity correlation data set for processing and application [4]. Big data described its characteristics mainly with "4V" features that 1) Volume: storing large, computationally intensive; 2) Variety: multi-Source, multi-format; 3) Velocity: growing faster, processing speed; 4) Value: scouring sand in the surf but precious [5-6]. According to the basic technical requirements of the WRDC, the logical structure of all nodes include five parts: resource layer, service layer, portal web, environment-guarantee and running environment, and on the basis of top layer design of the National Water Resources Informatization, main construction of WRDC includes infrastructure, data resources, information organization platform, environment-guarantee, etc[7-8]. As can be seen from the above description, the current node structure of the WRDC, mainly considers the processing of structured data, and some of the unstructured and semi-structured data, but cannot process WRDC with "4v" characteristics with conventional methods for adaptive analysis and design. Thus, to meet needs of supporting to process water resources big data, it is necessary to extend system architecture and infrastructure, more importantly, need to extend the technology concept of WRDC, so as to adapt to requirements of processing the water resources big data[9-12].

Aiming at the shortcomings of the traditional WRDC for processing water resources big data, this paper constructs a scene analysis model for water resources big data, focusing on the generation and application of digital scene on the condition of big data, and supporting the new mode and application of data center with causal analysis combining with correlation deduction, to rapidly and real-timely process and dynamically analyze water resources big data and solve the evaluation of water situation and other issues, for meeting the needs of all kinds of water conservancy decisions.

II. THE DEFINITION OF WATER RESOURCES SCENE

Water resources scene is a collection of information that describes the status of all kinds of water resources objects within some given area (range) and some given time. The formula employed is given as:

$$S(\mathrm{D},\mathrm{t}) = \{D \times t \times O, O \subset D, T \in [t - \Delta t, t] \wedge O(\mathrm{T}) \subset O\} \quad (1)$$

Where:

S: water resources scene;

D: some given area, which can be geographical scope, can also be the concerned water resources business areas, or the combination of both;

t: some given moment;

O: the water resources objects for some given area, and using the unique identification;

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.77

O (T): the state of water resources objects for some given moment;

Δ t: time increment.

The unchanged elements in object properties are defined as the elements of water resources scene framework within the concerned time to reduce data access number. The element collection of framework of water resources scene (a scene frame) can be used as a kind of special objects for encoding for assignment, storage and exploit repeatedly in the WRDC. The base map of water resources digital map is one of the typical scene frameworks.

III. THE OPERATION OF WATER RESOURCES SCENE

The operation of water resources scene mainly includes:

1. Scene generation: According to the specified area and the given time, all or part of the objects in the scene may accomplish the relevant information once extraction and integration by identifying objects in the WRDC. The usage of scene frameworks will effectively save computing resources.

2. Scene feature extraction: Optionally, choosing some or all attributes of the objects from the scene as the feature of the scene for characterization of the status and changes in the scene.

3. Scene deduction: The status of S (t +D, Δ t) is inferred by the state of S (D, t) deduces, or its inversion.

4. Scene similarity analysis: Defining the similarity through the scene features to find a similar scene.

5. Association analysis: The association evaluation between the objects in the scene, and finding object relation results in change or remain stable state in the scene.

6. Scene transformation: It is realized by reducing or enlarging the scene of the designated area, or increasing or decreasing the granularity of objects in the scene, or increasing or decreasing the objects in the scene.

7. Scene recurrence: It is realized forward or backward by changing specified time in order in the same scene.

8. Scene visualization: applications of various information visualization techniques to show the state of the scene.

Because the operation of water resources scene has to deal with massive structured and unstructured data, the WRDC needs to have the ability to quickly organize information by subjects, and support the processing of big data.

IV. THE CONSTRUCTION OF WATER RESOURCES SCENE

According to the definition of water resources scene, the construction of water resources scene is to choose the appropriate area (field) and time to find the objects within the region, determine the particle size of each object (basic objects or composite objects), comprehensive query and organize all attributes of the objects based on the metadata of objects, and form the water resources scene (data set) finally. Since this process may involve massive attribute data of objects (structured and unstructured) access (such as the regional scene analysis of watershed-level), requiring the WRDC to have the ability to process big data. Generation of water resources scene based on the WRDC is shown in Figure1.

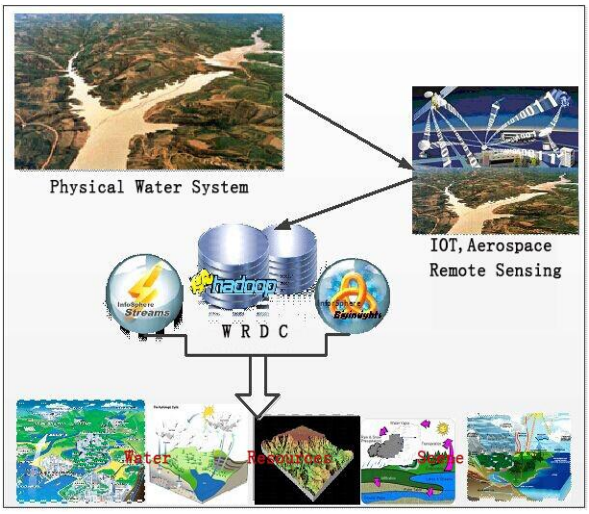


Fig.1 Generation of water resources scene based on the WRDC

The generate water resources scene according to the following steps:

1 . Choose the region or domain (problem domain), determine the scope of the object through the map or artificial operation, and assign some code to the selected region. This implies that the water resources scene can also be defined as a kind of compound of water resources object, or choose the defined framework as the water resources scene.

2. Find all water resources objects in the selected region. If the water resources scene is newly defined, the constant data of properties of all the objects can be defined as the framework, and save the structure of the framework. On the basis of the framework of the scene, defining all time-varying properties, all objects are provided with codes for identification, and organizing object properties, forming and saving the object relations of the water resources scene based on the EPC encoding system in the end.

3. Define one or more the given time and apply service platform of the WRDC for the data extraction function of routine data and big data, to extract scene data for scene analysis, based on the object relations of the water resources scene under the EPC encoding system.

V. THE METHODS OF WATER RESOURCES SCENE ANALYSIS

Water resources scene analysis, is essential association analysis for multi-objects. With analysis of the relations of attributes and methods belong to each object in the scene and the evolution of the different scenes to achieve the evaluation of water situation and other issues. The main techniques include:

1. Feature extraction for water resources scene: through the water resources scene analysis to accurately reflect the regional flood control situation, the condition of water resources supply and demand or decision-making for water function areas, for

identifying several key sensitive object properties, defined as the features of water resources scene, and simplifying the complexity of the scene analysis and improve efficiency.

2. Situation analysis and forecast for regional macroscopic water resources: through scene deduction to realize the reappearance of macroscopic water resources situation of different historical moments, and analysis for the main trend, for making predictive analysis for the future situation, such as the evolution analysis of the regional water ecology, prediction of drought development, etc.

3. Historical situation reference: through scenes features to define similarity and find similar historical scenes, for achieving specific reference to the historical experience of the current situation and the judgment of the development trend, the consequences of different decisions of flood dispatching prediction, etc.

4. Comprehensive association analysis: through the analysis of the relationship between the objects in the scene, and finding object relation results in change or remain stable state in the scene, so as to obtain the key objects and their interactions result in the development of the situation or unintended consequences, and provide a reference for the comprehensive decision-making.

5. Comprehensive application: the application of water resources scene analysis is different from the conventional computer application for water conservancy business, and the main differences lie in: water resources scene analysis is essential that according to resources of the WRDC such as computation, storage and software tools, in order to meet needs of water resources macroscopic decision-making, data analysis experts for water resources, (water resources experts, water resources analysts) define the corresponding scenes, and comprehensively apply scene operations, such as scene feature extraction, scene deduction, scene similarity analysis, multi-object association analysis, scene transformation, scene recurrence and visualization, for drawing the corresponding situation judgment for water resources, which is a typical man-machine coordinated way, and one of the important applications of the WRDC, and also the only way to make full of resources advantages of the WRDC.

Water resources scene analysis is not the information application of procedural approach, but the non-process-type "intuition" approach all over the territory. Therefore, in addition to develop information supporting functions in the WRDC, the cultivation and equipment of the water resources data analysts has become one of the development focus of the WRDC in the future.

Water resources scene analysis is not a single independent water conservancy business application system in the WRDC, but a part of the application service platform of the WRDC, closely combines with the EPC encoding, and achieves aims with the configuration and extension of the application service platform.

VI. THE PLATFORM SUPPORTING WATER RESOURCES BIG DATA

Similar to the platform supporting traditional data, the platform supporting big data mainly consists of a number of the commercial system software. After the analysis of solution to big data of global manufacturers, it can be seen that the basic support environment for big data processing, is mainly the current international popular Hadoop system, which has become the DE facto standard and supported MapReduce for batch and real-time processing based on big data organization of NoSQL such as HDFS[13], HBase[14], and supported data mining and real-time processing of big data.

Hadoop itself is open source software, on the basis of which, People may self-organize corresponding tools research and development, according to the application characteristics of water resources big data. But, if people don't choose business solution, they will have to do a lot of experiments, fully develop and extend the corresponding tools and applications, and the development workload is big, cycle is long. Considering demands of the WRDC supporting applications of big data, the IBM solutions of big data [15], may achieve the goal of the construction of WRDC supporting water resources big data processing.

According to the actual conditions of WRDC, the establishment of platform supporting big data, divided into two steps. First of all, determining a fundamental frame of supporting the big data for the WRDC, which extends basic architecture of the traditional WRDC, configures software supporting big data platform, establishes physical storage for big data, and imports the unstructured data into the big data storage system, to accomplish comprehensive test for data access. Secondly, all the data import the big data storage system, and realize big data analysis for all the data in WRDC.

In essence, big data storage mainly improves the efficiency of data storage and access for massive data. Big data analysis is the association analysis for large-scale complex dimensions, and the results of analysis are complicated association relationships, while the results of the traditional data processing are usually precise causal relationship.

Therefore, recent in the construction of WRDC, the main functional requirements of big data platform is to solve the problem of efficient storage for unstructured data, and be combined with structured data quickly for extraction and integration according to the objects, for supporting the implementation of all operations for the water resources scene.

Realizing object extraction and integration on unstructured data and structured data, with real-time analysis and batch processing on the IBM platform, so as to achieve applications of water resources scene combined with routine analysis processing. As all data from data center migrated to big data for storage, data processing will be greatly simplified. The basic data flow diagram of big data supporting water resources scene and routine analysis is shown in Figure 2.

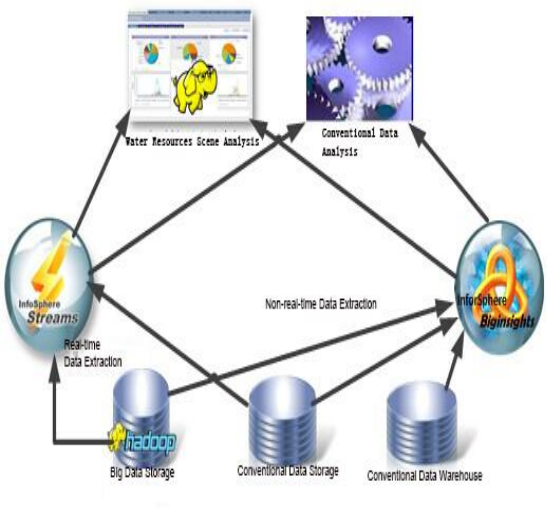


Fig.2 The basic data flow diagram of big data supporting water resources scene and routine analysis

VII. CONCLUSION

The construction of Smart Water puts forward higher requirements for the analysis and application of the water resources information. Typical application such as on-line analysis, digital simulation, mainly takes account of the causation analysis; while the massive association analysis based on the fourth paradigm for scientific research, is becoming the main stream under the condition of application of big data. As far as water resources information, according to the resources of WRDC, constructing the new mode of causal analysis combined with association analysis has become one of the important trends for application of the WRDC.

Unlike causal analysis, association analysis of the water resources information, mainly solve the problem of the evaluation of water situation and other issues, and the key lies in the construction and deduction of water resources scene.

ACKNOWLEDGMENT

This research was supported by the Major Program of the National Social Science Foundation of China under Grant No. 2012&ZD214; the Major Research Plan Training Project of the National Natural Science Foundation of China under Grant No. 90924027; the Key Project of the National Natural Science Foundation of China under Grant No.41030636; the Graduate Research and Innovation Projects in Jiangsu Province under Grant No. CXZZ13_0261.

REFERENCES

[1] Kumar, S., Peters-Lidard, C., Tian, Y., Reichle, R., Geiger, J., Alonge, C., ... & Houser, P. (2008). An integrated hydrologic modeling and data assimilation framework. Computer, 41(12), 52-59.

[2] Abbott, M., & Vojinovic, Z. (2009). Applications of numerical modelling in hydroinformatics. Journal of Hydroinformatics, 11(3-4), 308-319.

[3] Gourbesville, P. (2009). Data and hydroinformatics: new possibilities and challenges. Journal of Hydroinformatics, 11(3-4), 330-343.

[4] Big data [EB/OL].[2012-10-02}. http://en.wikipedia.org/wiki/Big_data.

[5] Science. Special online collection: Dealing with data [EB/OL]. [2012-10-02]. http://www.sciencemag.org/site/special/data/.

[6] Nature. Big Data [EB/OL].[2012-10-02].

[7] Information center of Ministry of water resources. The development of Chinese water Resources informatization"Twelfth Five Year Plan"[R].Beijing: Ministry of Water Resources Informatization Leading Group Office, 2012.

[8] Information center of Ministry of water resources. Top-level Design about Chinese Water Resources Informatization [R].Beijing: Ministry of Water Resources Informatization Leading Group Office, 2010.

[9] Barwick, H. The" four Vs" of Big Data. Implementing Information Infrastructure Symposium. 2012-10-02]. http://www. computerworld. com. au/article/396198/iiis_four_vs_big_data.

[10] Speitkamp, B., & Bichler, M. (2010). A mathematical programming approach for server consolidation problems in virtualized data centers. Services Computing, IEEE Transactions on, 3(4), 266-278.

[11] Bi, J., Zhu, Z., Tian, R., & Wang, Q. (2010, July). Dynamic provisioning modeling for virtualized multi-tier applications in cloud data center. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on (pp. 370-377). IEEE.

[12] Hill, R. R., Stinebaugh, J. A., & Briand, D. Wind Turbine Reliability: A Database and Analysis Approach.

[13] Apache Software Foundation. Apache HDFS.http://hadoop.apache.org/hdfs/.

[14] Hbase Development Team.Hbase: Bigtable-like structured storage for hadoop HDFS. 2009. http://wiki.apache.org/hadoop/Hbase.

[15] IBM Software solutions.http://www-01.ibm.com/software/

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

An improved PageRank algorithm based on web content

Zhou Hao, Pu Qiumei, Zhang Hong, Sha Zhihao
Institute of Information Engineering
Minzu University of China
Beijing 100081, P.R. China
muc_zhouhao@163.com

***Abstract*—With the development of Web technology and more kinds of information, how to provide high quality, relevant search results become a huge challenge to the current Web search engines. We analyze the shortcomings of PageRank algorithm and Weighted PageRank algorithms and make targeted improvements. By judging the relation between different webpages based on web content, the improved PageRank algorithm improved the user search precision.**

Keywords- PageRank; Weighted PageRank; web content; search precision

I. INTRODUCTION

With the rapid development of the network, the number of Internet users increased dramatically, according to the CNNIC "35th Statistical Report on Internet Development in China", the web users in China reached 649 million, Internet penetration rate was 47.9%, while Internet information growing exponentially. Faced with massive information and large-scale Internet users, the most important issue of how to make web users quickly and efficiently search for desired information is a high performed network search mechanism. [1] And one of the key technologies to solve this problem is Page Rank technology, because as a core part of the search engine, this technology is an important indicator of the level of search engine quality. [2][3]Currently, there are a lot of page ranking algorithm, including two categories: algorithm based on web pages content analysis and algorithm based on the link structure analysis. [4]The famous web search engine ranking algorithm used by Google is PageRank algorithm based on links structure analysis and the successful application of this algorithm confirming the practical application value and theoretical research value it has. Through the calculation of more than 500 million variables and 2 billion vocabulary, PageRank can make an objective assessment of the importance of web pages. PageRank does not count the number of direct links, but a link from page A to page B means page A cast a vote to page B. Thus, PageRank will assess the importance of the page according to the number of votes received by B.

In addition, PageRank also assess the importance of each vote page, because some pages are considered voting with high value, so that web pages get links from these web pages can receive a higher value. Important pages get high PageRank value, so that appears at the top of the search results. Google technology uses integrated information online feedback to determine the importance of a page. Search results without manual intervention or manipulation, which is why Google would become a source of information widely trusted by users, the impact is not paid placement and impartial.

II. RELATED WORK

A. PageRank algorithm

On the basis of traditional citation analysis, in 1998 two graduators Lawrence Page and Sergey Brin presents a Web-based link analysis algorithm at Stanford University, School of Computer Science. This algorithm is the oldest traditional PageRank algorithm. The algorithm uses the page link relationship between the structures of the entire network of relationships represented as a Web map.

PageRank calculations take full advantage of the two assumptions: proceed as follows:

(1) At the initial stage: the relationship built up through the links page Web map, each page set the same PageRank value, by calculating the number of rounds, it will be the final PageRank values for each page obtained. With calculation of each round of calculation, the PageRank value of current page will be continuously updated.

(2) Update the PageRank score of a page in a new round: in a new round, the calculations of updated PageRank score are as follows, current PageRank value of each page is distributed evenly to the chains contained within this page, so that each link get corresponding PageRank score. And each page will sum all chain weights to this page, then we get the new PageRank score. Each page completes a round of PageRank calculations as each page is given an updated PageRank value. [5]

If there is a link in page T connect to page A, it indicates that the owner of T thinks A is important, so that the part of the score given to page A is: PR (T) / L (T).

Where PR (T) is the PageRank value of T, L (T) is the number of chains of T, PageRank score of A is summed of a series of similar pages like T.

The number of votes a page gets is determined by its importance to all linked pages, a hyperlink to a page is one vote to the page. PageRank of a page is determined by the importance of all pages linked to it through the recursive algorithm. A page with more links have a higher rating, on

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.78

the contrary, if a page does not have any links, it is no hierarchy.[6]

PageRank algorithm sees the links to other pages as votes for the chain to the website, if a page has more links to other pages, the more authoritative this page is, so the page in search results is more forward. Algorithm formula is as follows:

$$PR(u) = (1-\alpha) + \alpha \sum_{v=B(u)} \frac{PR(v)}{Out(v)}$$

In the formula, α is called damping factor, usually ranging 0.85; B (u) is the collection of all pages that link to page M; Out (v) is out of the page. As we can be seen from the formula, PageRank algorithm webpage distributes PR (PageRank) value averagely to the forward link of each page, and page Ⅱ PR value is the summed value of all the pages that are linked to web Ⅱ. [7]

B. Weighted PageRank Algorithm

There are two types of links, income links and outcome links. While PageRank algorithm is based on the structure of the proposed links, but in-depth study will find that PageRank algorithm is based only on the structure of the page chain and distributes PR values equally. Thus, Xing, etc. expanded the traditional PageRank algorithm and proposed Weighted PageRank algorithm through further analysis of the link structure. The algorithm distributes PR based on income web links factor $W^{in}_{(v,u)}$ and outcome web link factor $W^{out}_{(v,u)}$, the formula is as follows:

$$PR(u) = (1-\alpha) + \alpha \sum_{v \in B(u)} PR(v) W^{out}_{(v,u)} W^{in}_{(v,u)}$$

B (u) is a collection of all backlinks of web u. Factors $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$ are calculated as follows:

$$W^{in}_{(v,u)} = I_u / \sum_{p \in F(v)} I_p$$

$$W^{out}_{(v,u)} = O_u / \sum_{p \in F(v)} O_p$$

F (v) is a collection of linked pages of v $W^{in}_{(v,u)}$ is based on income factor u and the sum of income v $\sum_{p \in F(v)} I_p$ of web v, results in the income factor of link (v, u); $W^{out}_{(v,u)}$ is based on the outcome of u and the sum of outcome $\sum_{p \in F(v)} O_p$ of web v, results in the outcome factor of link (v, u).[8]

Xing has shown that Weighted PageRank combined income factor and outcome factor of web gets better sort results than the traditional PageRank algorithm. But Weighted PageRank algorithm still only considers links structure, so it has some of the same drawbacks as the traditional PageRank algorithm. Including:

(1)Topic drift. Since there is no consideration of web content and related topics, therefore it can not recognize whether the search results are relevant to the search topic, causing some searched pages off-topic.

(2)Favors the old pages. Compared to the old page, a new page exists a short time, links relation was simple and PR value obtained was small, sort position was lower. Obviously, it is unreasonable to hot topic for the new page.

(3) Ignore user interest. Based only on the link relationship regardless of the user's interest in a web page, is also drawback of Weighted PageRank Algorithm.

III. IMPROVED PAGE RANKING ALGORITHM

Searching for a particular field, we believe that these algorithms lack of consideration for the relevance of pages, without consideration of web content and search related topics, the PR value a page distributes to outcome pages should be concerned with the relevance. We analyze HTML text of each page, it is easy to get <anchor> and <title> tag content, and can get specialized vocabulary and terminology of a particular field, if the level of the page <anchor> and the next level of < title> tag contents and any two of the three keywords matches, we can think that this link is valid, this page can get distributed PR value, if <anchor> tag and the <title> tag content and keywords do not match, this is an invalid link, the next level page cannot get the PR value from this page, this can make the PR value distribution of the site more reasonable.

At the same time, we think that the more valid links a web has, the more reliable the web is.

We define the following calculation:

$$PR(\mathrm{u}) = (1-\alpha) + \alpha \left(\sum_{v \in P(u)} PR(v) * I_{(v,n)} \right) * Q_u$$

α is the damping factor, usually taken 0.85, P (u) is the set of all pages linking to u, $I_{(v,n)}$ *is the* PR value web u obtained from the web page v.

$$I_{(v,n)} = \frac{\beta_v}{n_v}$$

β_v is the number of valid links on the page, n_v is the total number of links on the page v and $I_{(v,n)}$ represent a valid

link proportion of the total number of links, namely: the PR value of v are allocated to each valid link evenly. After U gets page PR value from all page links, this value is multiplied by the credibility of the pageQ_u.

$$Q_u = \frac{\beta_u}{n_u}$$

β_u is number of valid links on the page, n_u is the total number of links on the page u, Q_u represents a valid link proportion of the total number of links, the credibility of the web page.

IV. EXPERIMENTAL EVALUATION

In the experiment, we selected the keywords in different areas:

1. We selected the keyword "sports", we get the following statistics:

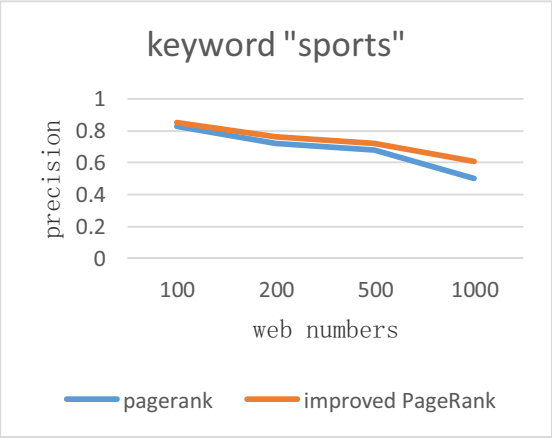


Figure 1. search result of keyword"sprots"

2. keyword "study"

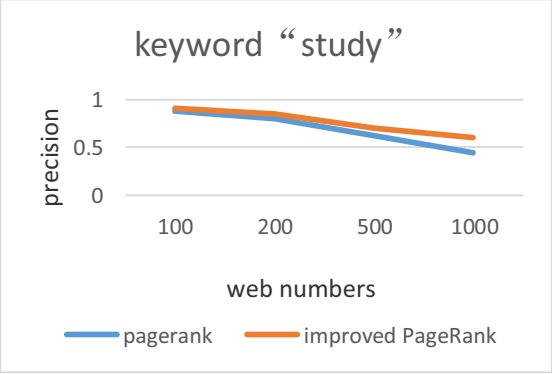


Figure 2. search result of keyword"study"

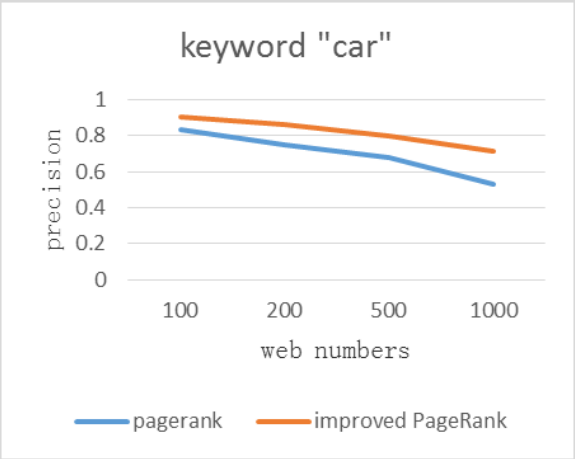


Figure 3. search result of keyword"car"

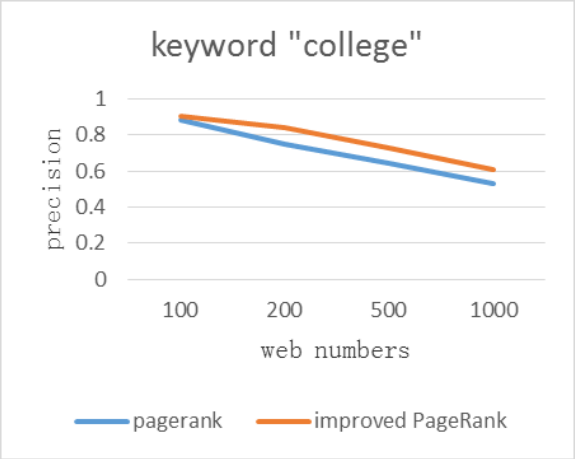


Figure 4. search result of keyword"college"

In this improved algorithm, we fully consider the credibility of links and relevance between pages in our experimental. The improved search accuracy can be seen in our statistics, which enables us to get more accurate search result in some specific fields. In the latter experiment, we will consider user feedback into account to the page, to enhance the search results for the user's satisfaction.

REFERENCES

[1] Kale M.; Thilagam, PS DYNA-RANK: Efficient Calculation and Updation of PageRank 2008

[2] Zhang Ji-Lin; Ren Yong- jian; Zhang Wei; Xu Xiang-Hua; Wan Jian; Weng Yu Webs ranking model based on pagerank algorithm 2010

[3] Al- Saffar, S .; Heileman, G. Experimental Bounds on the Usefulness of Personalized and Topic-Sensitive PageRank 2007

[4] Haveliwala, TH Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search 200 3

[5] Zhang Kun; Li Peipei; Zhu Baoping; Hu Manyu Evaluation Method for Node Importance in Directed-Weighted Complex Networks Based on PageRank 2013
[6] CAO Shan- shan; WANG Chong Improved PageRank Algorithm Based on Links and User Feedback 2015
[7] WANG Deguang; ZHOU Zhigang; LIANG Xu Analysis of PageRank Algorithm and Its Improvement 2011
[8] DUAN Huai-chuan; HU Ping Improved PageRank algorithm based on topic character and time factor 2010

Microblog sentiment analysis algorithm research and implementation based on classification

Yanxia Yang
Faculty of Information Engineering,
City College Wuhan University of Science and Technology,
Wuhan, China
yxy_job@163.com

Fengli Zhou
Faculty of Information Engineering,
City College Wuhan University of Science and Technology,
Wuhan, China
thinkview@163.com

***Abstract*—Under the background of today's information age, microblog obtains a rapid development. With the news on the microblog updating, in order to avoid the users getting lost in the ocean of information, emotion analysis of the information becomes urgent and important. This paper based on the implementation of microblog emotion mining of Bayesian classifier and SVM classification algorithm, making comparison through the analysis of the experimental results in processing speed and accuracy, has a reference value.**

Keywords- Bayesian; SVM; words segmentation; sentiment analysis

I. INTRODUCTION

As a platform combined information and social contact, the microblog platform attracts a lot of users to analyze emotion and study transference through its unique charm. Emotion analysis is mainly to mine views expressed by users from the text and the polarity of the emotion, that is, the positive, negative and neutral of the expression of a micro blog message is judged[1]. At present, the research is still in the initial stage, so this paper from the micro blog users' information, and the research is based on classification algorithm of the emotion of microblog author mining system implementation.

II. RELATED WORK

A. Pretreatment

Preprocessing text is to segment the text, and only through segmenting the text the characteristics of the text can be judged, analyzed and classified. Hence, reasonable partition Chinese character string is requisite[2].

Adopting Lucene Chinese word segmentation can realize the extraction of the Chinese word segmentation and keyword. Stop words, also known as function words, usually are some prepositions, pronouns and empty words, which have nothing to do with emotion in general texts. Since the majority of the microblog platforms support texts, pictures, expressions, audio, video, etc, it also needs to deal with some irrelevant symbols that have no practical significance in the emotion analysis and research.

B. Feature selection

After the text segmentation, what keywords are mainly considered as text features: word frequency, regional location and factors of segmentation of distance and position.

Considering the above factors, the weight formula of the selected candidate words is as below:

$$\text{weight}_i = \alpha \times \text{tf}_i + \beta \times \text{loc}_i + \gamma \times dis_i \quad (1)$$

In the formula(1), $weight_i$ is the weight of the candidate words $word_i$; tf_i is the frequency factor; loc_i is the location factor; dis_i is the distance order factor; α, β, γ are regulators of three factors.

The frequency factor uses the formula $\text{tfi} = \ln f_i$ (f_i is the candidate word frequency in the text) to calculate. The positions of each word in the text are recorded, and the

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.79

words are labeled in different positions of the text. If a word appears frequently in the text, the choice of its position is selected at the top of the position. In the process of text processing, the weight of words can be calculated, and the formula is shown as below:

$$loc_i = (w_i - 1) / (w_i + 1) \quad (2)$$

In the formula(2), W_i is the position value that candidate words is labeled in participles.Through experiment, this paper mainly get through a linear function: $val_i = a \times i + b$ is to mark the distance order value of word segmentation (i shows the order that the words appear in the text, while a and b are adjustable constant factors). Then, calculating the distance order weight via the following formula:

$$dis_i = val_i / \ln val_i \quad (3)$$

In the formula(3), val_i refers to the distance from the position of the word segmentation, in which first appears, to the beginning of the text. By introducing the logarithmic function into the sub formula, the nonlinear features of the feature term can be better characterized.

2) Weight factor training

This paper adopted the least mean square error training formula to train the adjustment factor of formula.

The first step is to set the value of the adjustment factor, then record the weight of each word in the text by calculating, and rank the results according to the weight from high to low. Supposing that the words collection of the i text after the k times calculation of weight and sorting of words collection is $V(k,i)$, and the text of the training words sort is recorded as V_i. According to the difference of weight sorting of words between the $V(k,i)$ and V_i. set the sorting difference:

$$diff = \sum_{j=1}^{n} (sort_{(i,j)} - sort_{(k,i,j)}) \quad (4)$$

In the above formula(4) is the j participle in training sort set and the sort order of test sort concentration after the k calculation in $sort_{(i,j)}$ and $sort_{(k,i,j)}$ texts.

Then adjust the value of each adjustment factor (α, β, γ) by the next formula:

$$w = w + \eta \times diff \times sec \quad (5)$$

In the formula(5), W is adjustment factor, $diff$ is a very little constant factor, sec is the value of the current test factor.

C. Text classification algorithm

Bayesian classification algorithm, is a algorithm which uses knowledge of mathematical probability to classify. It is able to applied to many aspects, is a easy method with high accuracy of classification and speed.

Bayesian classification is divided into three stages[3]:

The first stage is the preparation stage. This stage is mainly to input all the data to be classified, determine the feature attributes and output training samples after a series of processing.

The second stage is training stage of classifier. The task of this stage is to generate separator, calculate the frequency of each category appeared in the training samples and conditional probability estimation of each feature partition for each category, record the results and identify the category according to the classification results.

The third stage is application stage. The task of this stage is to classify the items to be classified by using classifier, carry on the first two stages and finish the final classification. This stage is mainly a integration of the first two steps, classifying the category that the sample belongs to through the processing of the first two steps and realize the classification.

Support vector machine (SVM) is based on VC dimension theory of statistics theory and the basis of structure risk minimization principle, which seeks the best comprise between the model complexity and learning ability according to the limited sample information to obtain the best generalization ability[4]. The basic idea of SVM is：make a mapping of nonlinear support vector machine input a higher dimensional space and turn it into linear support vector machine, then select optimal linear classification surface in this new space, and this transformation is usually achieved by using properly defined inner product function(kernel function), which is based on support vector in training set[5].

D. Construction of emotional vocabulary Ontology

At this stage, the task is to collect the core concepts

and relationships that can express people's opinions. Constructing emotional vocabulary ontology is for a more complete expression of semantic information that the emotional vocabulary contains, such as the similarity between the sentiment orientation and vocabulary of words, turning and progressive relationship etc, and also is convenient for organizing and sharing emotional words, thus providing forceful analysis basis for public sentiment analysis research[6]. This paper mainly selected core vocabulary of emotion analysis words set published by China HowNet to act as information source of constructing dictionary[7].

Negative words, adverbs of degree, and the logical conjunctions of turning and progression can have an effect on the tendency of the subjective sentence. This paper also set up the negative words, adverbs of degree and conjunction sets[8]. Based on the negative words, degree adverbs and conjunction sets released by HowNet, there includes 18 negative adverbs, 40 conjunctions and 188 adverbs of degree[9].

III. System implementation

A *Naive Bayes subsystem*

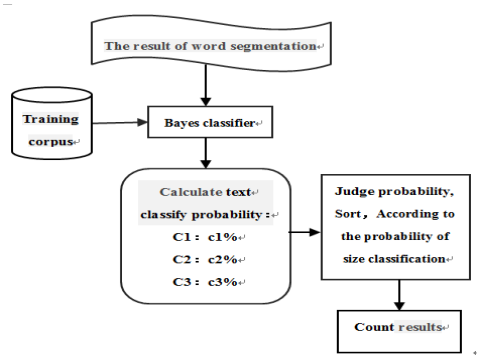


Figure 1. Flow chart of Bayesian classifier

B *SVM classification subsystem*

Implementation process of SVM classification subsystem is shown in Figure 2:

IV. Experimental results

A *Experimental data*

Test data contained two parts. There are 998 sentiment classification data used for training and calculating probability of data to be measured. The data to be measured contained 6626 microblog data, and the classified statistical work is finally finished through the naive Bayesian classifier and the Libsvm classifier.

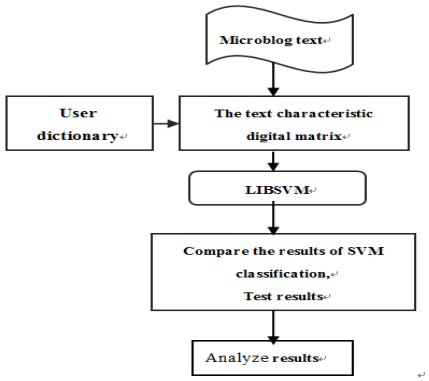


Figure 2. Implementation process of SVM classification

B *Bayesian model test results*

According to Bayesian algorithm, it requires the treatment of training corpora containing three emotion categories, including positive emotion, neutral emotion and negative emotion.

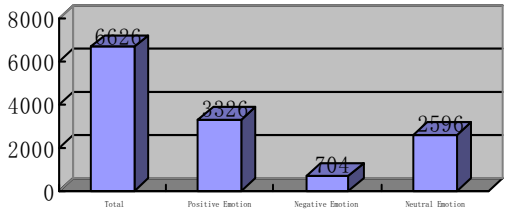


Figure 3. Results of Bayesian text classification

C *SVM model test results*

For the SVM algorithm, because the machine cannot recognize the Chinese text, transforming into digital matrix is needed, and then turn test text into matrix to be tested which contains emotional attributes. The text classification results implemented by SVM algorithm include output iteration number, minimum value solved through quadratic programming (optimal value of SVM issues), constant project literal *b* of solution as well as the training data number, support vector number and the final results of

model accuracy alignment. The text classification was realized through formula $g(x) = wx + b$:

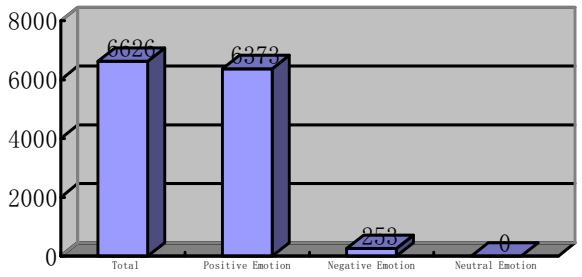


Figure 3. Rresults based on SVM sentiment classification

The paper selects the same test text, and the results based on Bayesian and SVM algorithm are different from this paper's. There are three types, that is, positive emotion, negative emotion and neutral emotion, in the results of text classification while using Bayesian algorithm. As is shown in Figure 3, positive emotion and neutral emotion account for a larger proportion in the text; when using SVM algorithm to realize text classification, shown in Figure 4, as to neutral emotion, this paper classified it as positive emotion, but negative emotion accounts for a smaller proportion in both, which led to a big difference between the results realized by two kinds of algorithm.

V. Concluding remarks

In this paper, Bayesian algorithm and SVM, used to realize the text classification respectively, after comparison, turn out to have advantages and disadvantages each: Bayesian algorithm implementation process needs to spend a lot of time processing large amount of text information and is not fast enough; SVM algorithm is relatively fast in processing，and the reason is that the key of the algorithm is to find reasonable function, which can transfer texts into needed text feature digital matrix. What's more, the accuracy of the function determines the accuracy of the results because its data processing doesn't require a large number of segmentation comparison. Therefore, SVM has a better effect while dealing with plenty of texts.

ACKNOWLEDGMENT

The work in this paper is partially supported by the Education Department Foundation of Hubei Province of China under Grant No. B2015360 and Wuhan University of Science and Technology city college scientific research project under Grant No. 2014CYYBKY011.

REFERENCES

[1] XIE Lixing, ZHOU Ming, SUN Maosong, "Hierarchical Structure Based Hybrid Approach to Sentiment Analysis of Chinese MicroBlog and Its Feature Extraction"[J], Journal of Chinese Information Processing, 2012,26(1)pp:73-83.

[2] Kun Zhou, "Research of Chinese Text Classification based on Improved Vector Space Model"[D],Beijing Institute of Technology, January,2015.

[3] ZHANG Xin, MA Yong, CAO Peng, "Traffic Identifying Method for Trojan Detection upon Bayesian Classification Algorithm"[J], Netinfo Security, 2012.(8)pp:115-117.

[4] YANG Bin, LU You, "Classification Method of Support Vector Machine Based on Statistical Learning Theory"[J], Computer Technology and Development, 2006,(11)pp:56-58.

[5] MA Jinna, TIAN Dagang, "Research on Chinese-text Automatic Classification Based on SVM"[J], Computer and Modernization, 2006,(8)pp:5-8.

[6] K.T. Durant , M.D. Smith, "Mining Sentiment Classification from Political Web Logs"[C],In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2012),August, 2012.

[7] Subject sentiment analysis research Based on ontology, http://www.docin.com/p-976811905.html, [EB/OL］.

[8] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins. "Information Diffusion Through Blogspace" [C], In Proceedings of the 13th International Conference on World WideWeb, 2013 pp: 491-501.

[9] WANG Xiaodong, WANG Juan, ZHANG Zheng, " Computation on orientation for subjective sentence based on sentiment words ontology" [J], Journal of Computer Applications, 2012,32(6)pp：1678 -1681.

Opinion Leaders Discovering in Social Networks Based on Complex Network and DBSCAN Cluster

Xiaoli Lin
Information and Engineering Department of City College,
Wuhan University of Science and Technology,
Wuhan, China, 430083
aneya@163.com

Wei Han
Information and Engineering Department of City College,
Wuhan University of Science and Technology,
Wuhan, China, 430083
478178271@qq.com

***Abstract*—The opinion leaders play an important role in the process of network public opinion spreading. In order to quickly and efficiently discover the opinion leaders, this paper analyzes the characteristics of complex networks in social networks and proposes density-based spatial clustering of applications with noise algorithm based on local community detection method. With Sina micro-blog user as the research object, the feature vectors of opinion leaders are extracted as the training set, then the characteristic means of the subclass are obtained, from which the user groups with the community opinion leader characteristics can been identified. Finally, DBSCAN algorithm is compared with the K-means algorithm and the average path length difference algorithm by using the same data set. The experiment results show that DBSCAN algorithm can be more accurate and more effective to find community opinion leaders.**

Keywords-DBSCAN cluster; SNS; complex network; opinion leader

I. INTRODUCTION *(HEADING 1)*

With the rapid development of Internet, Social Networking Service (SNS) has become an increasingly important communication platform [1]. On this platform, people always tend to communicate in different groups, so there is an amount of information which shows the relationship between each group. For example, a group of people are better active than others and have great influence, which means those ones are more likely in the dominate positions. So they are able to obtain information from media or other ways to publish their opinions, which is defined the opinion leader. Many researches show that the opinion learder pays a significant role in expressing ideas, transmitting information, leading thinking and coordinating social issues. As a result, it can control social consensus. Therefore, mining opinion leaders of SNS has become a heated point. Also, it has become the hot topic to mine the relation between users with the same hobbies.

A number of research works about social networks services have been done [1,2,3] such as twitter, but little is about social networks services in China. Sina Micro-Blog is the first social networks service in China and is growing fast in recent years. The length of a text message posted by users on Micro-Blog is limited to 140 characters so users can post their messages lightheartedly. Sina Micro-Blog provides a function for gathering the messages that are posted by people who are expected to provide information which are useful for a user.

In this paper, Sina Micro-Blog platform as the research object, we can automatically obtain the authorization and grab the micro-blog data from the application interface by using Python language and Web automation tools. Then, the data is converted into the required format for loading. After processing the data, each user of Micro-Blog is defined as node in the graph, and we extract the in-degree, the out-degree, the number of posts, the reply number and other information. Then, the community user feature vector is established. In addition, this paper proposes the improved DBSCAN algorithm to analyze the opinion leaders in social networks based on complex network.

This paper consists of the follows sections. Section 2 gives the relationship of complex networks and social network. Section 3 describes the data gather of Sina Micro-Blog user. Section 4 describes an improved DBSCAN algorithm. Section 5 shows some experimental results and discusses the effectiveness of the method. Section 6 describes conclusion and future directions.

II. COMPLEX NETWORKS AND SOCIAL NETWORK

In the social network, each user can be as nodes of a graph, and edges between two nodes can be as the relationship of users.

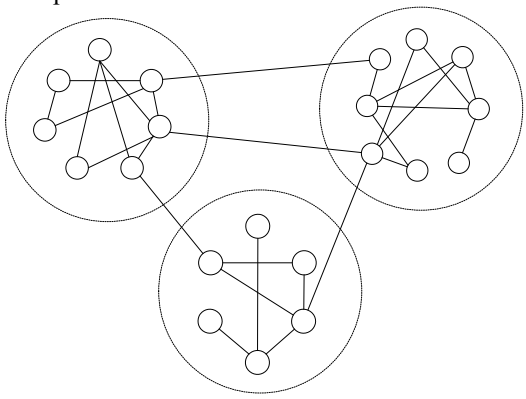


Figure 1. Social network user relationship network

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.80

The more frequent the interaction between the users is, the closer the edges between nodes are. The closely relationship may hidden potential relationship, such as interest association et al. [4], which constitutes a complex social relationship network. The complex network [5] is a network that has the characteristics like part or all self-organization, self-similar, attractor, small-world, scale-free, as shown in Figure 1.

III. SINA MICRO-BLOG DATA GATHER

A. OAuth2 Certification OF Sina Micro-Blog

OAuth is the international open standard for secure API authentication. OAuth is a token-passing mechanism that allows users to control which applications have access to their data without revealing their passwords or other credentials in the third application. The operation of validation and authorization are simpler and safer based on OAuth2.0.

When a user first tries to perform an action that requires API authentication, direct the user to Sina Micro-Blog authorization server. If the user consents to authorize your application to access those resources, Sina Micro-Blog will return a token to your application. Depending on your application's type, it will either validate the token or exchange it for a different type of token. For example, a server-side web application would exchange the returned token for an access token and a refresh token. The access token would let the application authorize requests on the user's behalf, and the refresh token would let the application retrieve a new access token when the original access token expires.

B. Authorized

After getting App_key and App_secret, the authorize code of table 1 is executed .Then the authorization Webpage will appear.

TABLE I. THE AUTHORIZE CODE BASED ON PYTHON

```
import sys
import weibo
import webbrowser
APP_KEY = '*********'
MY_APP_SECRET = '******'
REDIRECT_URL = 'https://api.weibo.com/oauth2/default.html'
api= weibo.APIClient(APP_KEY, MY_APP_SECRET)
authorize_url = api.get_authorize_url(REDIRECT_URL)
print(authorize_url)
webbrowser.open_new(authorize_url)
code = raw_input()
request=api.request_access_token(code,REDIRECT_ URL)
access_token = request.access_token
expires_in = request.expires_in
api.set_access_token(access_token, expires_in)
print(api.statuses__public_timeline())
```

C. Micro-Blog Data Acquirement

The process of grabbing automatically user information is illustrated as followed figure 3, which has been research in [6]. The system will start to analyze the target user's information after user ID was input. Each friend uses' information of the target user will be visited by the system automatically. If the friend user's information has been saved, the system will jump to another one. An inner counter is used to store the users who have been visited, which can help the system to continue the visit operation to the next friend user till all friend users are visited.

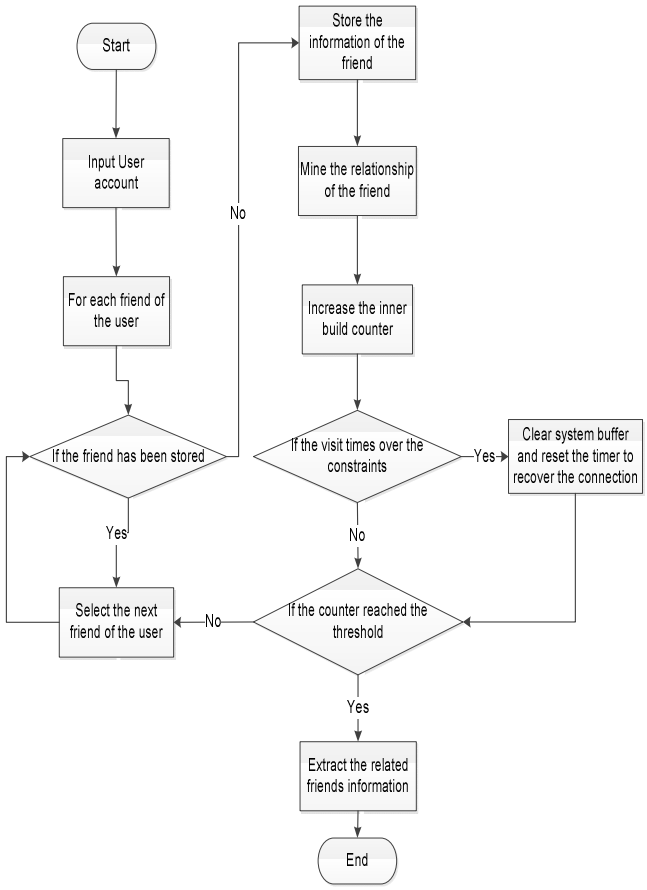


Figure 2. The Process of Grabbing User Information

IV. DBSCAN ALGORITHM

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester [7]. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together [8].

DBSCAN does not need to define the number of clusters, and it can handle discretional shaped clusters and can even search a cluster absolutely surrounded by other different cluster. The determination of two parameters Eps and MinPts is very key and even sometimes sensitive to the effectiveness

of clustering for the reduced set. The table 2 lists the iterative process.

TABLE II. DBSCAN ALGORITHM

Input: The information of Micro-Blog user
Output: The sets of cluster

Begin

Initialization: Place to unvisited state for all objects in data set.

Visit every object in data set and execute following operations:

If it belongs to a cluster then

Continue;

Else

1. Place to visited state for the current object;
2. Generate the neighborhood of the object

If the nodes in the neighborhood less than the threshold then

Remove the object from data set to noise set;

Else

1) Generate a new cluster with current object as it's center point;
2) Add all nodes in the neighborhood to the cluster;
3) Visit the nodes in the cluster, generate neighborhood of each node and create new cluster depending on nodes in the neighborhood

End

V. EXPERIMENTS

DBSCAN clustering algorithm has been realized by Python. The six difference clusters are obtained after the data sets are trained and tested. It can be seen from table 3 that the user numbers of the cluster 3, the cluster 4 and cluster 5 are fewer. The characteristic mean values corresponding to the cluster are shown as table 4. The cluster 5 is sub class of the least members and the largest characteristic value.

In addition, the number of replies and number of forwarding comments are significantly higher than those of other clusters, and the reply length is greater than the length of forwarding comments.

These show that cluster 5 has a strong influence, which has the most suitable for the characteristics of community opinion leaders. The members are: Patricklee, gogoboi. Next is the cluster 3, the member number is fewer and the characteristic value is big, the members are: Ara_kimbo, rhea140, Lvana.

TABLE III. THE RESULTS OF USER CLUSTER BASED ON DBSCAN

Cluster ID	User Number	Proportion
1	65	1.4753%
2	3436	77.9846%
3	3	0.0681%
4	6	0.1362%
5	2	0.0454%
6	894	20.2905%

TABLE IV. THE CHARACTERISTIC MEAN OF USER CLUSTER BASED ON DBSCAN

Cluster ID	Number of replies	Number of Forwarding comments	D-value
1	123	145	22
2	45	12	-33
3	510	598	88
4	234	67	-167
5	1098	1320	222
6	56	77	21

In order to verify the performance of DBSCAN clustering algorithm, K-means clustering algorithm is also used to cluster the same data, and the experimental results are shown in Table 5.

TABLE V. THE RESULTS OF USER CLUSTER BASED ON K-MEANS

Cluster ID	User Number	Number of replies	Number of Forwarding comments
1	12	113	234
2	2546	65	342
3	3	898	1123
4	6	145	345
5	494	580	52
6	1345	265	123

It can be seen from table 5, the number of clusters 3 is fewer and the characteristic mean value is larger, and its users are: Levi, EagleWan, Lvana.

But through the observation, it can be found that EagleWan of the cluster 3 posts more, but never gets over other people's reply, without opinion leader's influence and

appeal. The user Levi posts less and responses less, yet not opinion leaders. On the meanwhile, opinion leader Patricklee and Gogoboi which selected by algorithm DBSCAN are the hottest in the community and with great influence.

In order to further verify the effectiveness of our algorithm, the APLD (Average Path Length Difference) discovery algorithm is used to obtain the opinion leaders with the same data set. The rank of the top five opinion leaders is shown as table 6.

TABLE VI. THE RESULTS OF BASED ON AVERAGE PATH LENGTH DIFFERENCE DISCOVERY ALGORITHM

Order	User	APLD
1	Patricklee	0.8976
2	gogoboi	0.5643
3	Lvana	0.5321
4	AneyaE	0.4231
5	rhea140	0.1856

Table 6 shows top five results calculated by APLD. There is only the user AneyaE is different from the result calculated by DBSCAN. With detail analysis, we can find out that the user AneyaE never post and seldom reply, although he has received others reply but he still can't seem as the opinion leader. In general, APLD and DBSCAN algorithm can both find out the opinion leaders of the community, but DBSCAN's accuracy is better than the APLD.

VI. CONCLUSION

In this paper, we propose a method to analyze the opinion leaders based on the information of Sina Micro-Blog users. According to the user information, a user correlation matrix is generated. In addition, the improved DBSCAN algorithm is used for topic clustering analysis based the information of Micro-Blog users. Experimental results demonstrate that the proposed method outperforms more accurate and effective. And the further research and explore will be needed. We also will optimize our method to find out the similarity of attention behavior in our future work.

ACKNOWLEDGMENT

The work in this paper is partially supported by the Education Department Foundation of HuBei Province of China under Grant No. B2013258, and in part supported by Wuhan University of Science and Technology city college scientific research project (No.2014CYYBKY011).

REFERENCES

[1] Kwak H, Lee C, Prak H and Moon S. 2010. "What is Twitter, a Social Network or a News Media?" In: the International World Wide Web Conference Committee (IW3C2). ACM, New York, 591-600.

[2] Benevenut F. 2010. "Characterizing user behavior in online social networks," In: Proceedings of ACM SIGCOMM internet measurement conference. ACM, New York, 4962.

[3] Weng J, Lim E-P, Jiang J and He Q. 2010. "Twitter rank: finding topic-sensitive influential twitterers," In: Proc. of the third ACM international conference on Web search and data mining, New York, NY, USA, ACM, 261-270.

[4] Saifan Wang, Fang Dai and Bo Liang. 2011. "A path-based clustering algorithm of partition," In: Information and Control, 40(1): 141-144.

[5] Jiawei Wu, Xiongfei Li, Tao Sun and Wei Li. 2010. "A density-based clustering algorithm concerning neighborhood balance," In: Journal of Computer Research and Development, 47(6): 1044-1052.

[6] Xiaoli Lin, Yanxia Yang. 2014."User Relationship Mining on Micro-Blog Based on K-Means Algorithm," 2014 International Conference on Computational and Information Sciences, 491-496.

[7] S Akbar，MNA Khan. 2014. "Critical Analysis of Density-based Spatial Clustering of Applications with Noise (DBSCAN) Techniques," International Journal of Database Theory & Applictaion, 7(1):p17.

[8] EN Nasibov, G ULutagay. 2010. "Comparative Clustering Analysis Of Bispectral Index Series Of Brain Activity," Expert Systems with Application: An International Journal. 37(3): 2495-2504.

Data Analysis of Distributed Application Platform Based on the R which Apply to Digital Library

Ningbo Wu
College of Information
Guizhou University of Finance and Economics
Guiyang, China
hn_dragon@163.com

Fan Yang
Library
Guizhou University of Finance and Economics
Guiyang, China
Ginger318@gmail.com

***Abstract*—Digital Library large data resource lack of analysis and use, in order to mining the value of big data resources, proposed platformization analysis and processing mode. By integrate R and Hadoop to construct distributed data analysis platform, many big data analytical can be decomposed into "large" and "small" data processing section, overcome before scheme puzzle on analytical of large dataset, improve the performance of data analysis, platform able to handle data analysis tasks.**

Keywords-component; big data; distributed;data analysis

I. INTRODUCTION

Big data era, Digital Library(DL) collections nine kinds of native big data resources. Including digital photos, digital documents, web archiving, digital manuscripts, electronic records, static data sets, dynamic data, digital art and digital media publications[1] .This digital resource have diversified data features and low-value density characteristics. Due to various constraints, own at the space but lack of analysis and use, this bring about DL long-term in the ivory tower. Ricardo is success integrate R and Hadoop to construct an extreme analytics platform project at the IBM, with R language powerful statistical analysis and processing capacity integrate Hadoop advantages of distributed processing on the large dataset, this combination have advantages of flexible data storage, query and depth data analysis[2] .In this paper we will present a new idea, use R+Hadoop to building a distributed computing and analysis platform, this platform can manage the digital resource in DL,purpose to improve data more open under the Open Access(OA) Development Model in DL. Activating the flow of knowledge state to meet the research for the library and learning needs, It can provides us a real-time processing and analysis.

II. BIG DATA ANALYZED USING DEMAND IN DL

In the future web3.0 era, the semantic web and linked data technologies innovation,this will bring a new round of opportunities and challenges for DL. The DL based on semantic web needs to research the link between metadata knowledge and resource information, associate related between resource information[3]. This means DL faced with the requirement of information recall and precision in big data era. Under the background of big data, DL usher in the demand of digital resources development and utilization of the deep-analytics, mainly performance in the following two aspects.

A. Platform Service Needs

McKinsey issued an open data: the flow of information release innovation capability in October 2013,this report pointed out that the open data for the global open education, transportation and energy, about seven domains adding value to $ 3 trillion[4]. Mobility and more open in the active data under the new requirements of data management, DL needs to use new techniques ,combine proactively push consumption patterns with passive consumption patterns to activate the flow of knowledge and information carrier. DL faced with the demand of open and share digital resource, this need establishment of a comprehensive data management and monitoring system integrate DL business chain, and implements data resource flows continue to add value in the process[5]. As shown in Figure 1 construct a public service platform with resource sharing and open access.

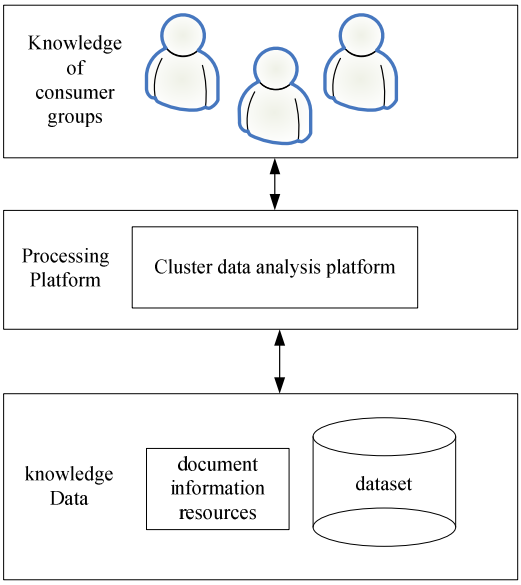


Figure 1. The platform level model for resource analysis

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.81

User demand for diverse information services, It requires a DL providing access to resources utilization ways, expect the platform can use the digital resource as the base, Integrate data manage , monitoring and analytics as one. With platform way to provide information services for users, under interactive mode information achieve the flow of the digital resource, explore the value of information resources.[6]

B. Integration resources and knowledge dscovery Needs

In our country, the DL Construct have achieve remarkable and have been able to rely on the structured information resources to provide users convenient access to information and knowledge retrieval services. But, on the other hand from the essence analysis ,current DL information services mode just as the resources digital store, search, lacking of use digital resource for deep development and semantic content analysis capabilities, original big data resource analyzed using in deep level still blank[7]. With the scientific research paradigm change in data-intensive scientific fourth paradigm, Data-driven scientific and academic exchange mode changes bring to the DL a new demand that large data should be analyzed and use frequently[8]. Under the new situation, DL should be embedded scientific research, set subject service as a whole.

In the knowledge center of big data era, DL towards a space-based collections to scientific research and academic exchanges third space transformation ,set manage resource and information services for a double feature. Facing with the needs of scientific research and knowledge development, DL should improve the ability to deal with the digital resource for deep-development and analysis in big data background, explore a new information services mode to solve the problem that distinguish from the tradition library[9]. The needs of scientific research ask for DL integrate original big data resource, Journal articles, achievements of scientific and technological ,etc, provide faceted clustering ,sorting and other related citation. At the same time, be able to make a unique identification for various types of knowledge objects (including the subject, field, academics, etc.), particle size analysis and related knowledge presented, meet researchers from different perspectives dig related resource information what hiding behind knowledge[10].

III. CONSTRUCT DISTRIBUTED DATA ANALYSIS PLATFORM

DL establish a depth analysis and make decision at the core of big data strategic thinking, Enhance the original big data resource intelligence analysis and processing capacity[11]. Integrate open scientific research data and the services that subject knowledge, provide real-time data analysis and visualization results published, making evidence analysis for scientific research in the field and frontier disciplines.

A. Platform Expected to Complete the Goal

Existing big data resources on the basis of the DL adopt science, open , heterogeneous, transparent, cross-platform and intelligent etc, strategies to build data analysis platform. Expect the platform achievement the following target.

1. Resource association found. Platform data analysis and processing module through R language that the advantages of statistics and mining with the text digital resource, implement semantic text mining. By using a unified standard metadata description, easy to find related resources[12]. To meet users semantic search services above the platform, overcoming the traditional model of the information can't recall problems.

2. Distributed storage resources. The platform through adopt Hadoop Distributed file resource storage strategy, implement resources long-term preservation .

3. Dynamic data analysis. Platform provides default scientific computing model, custom statistical analysis index system, set data released correlation factor, etc. Make full use of statistical computing R language and the advantage of MapReduce computing . The transition from level to reveal and describe data to data mining and knowledge discovery. Using regression analysis to fit the model formula for dynamic data, predictive analysis of data trends.

B. Platform System Model

Platform integrated DL's big data resources, adopt R + Hadoop technology integration process analysis is the core. The following figure 2 shows platform system constituted.

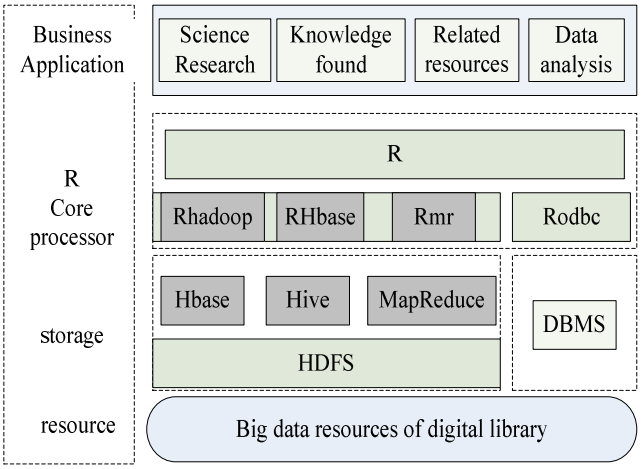


Figure 2. Platform System Model

The platform system mode include four levels.The lowest level is the big data resources, for resources to take structured and unstructured two storage, R core processor layer through R language packages Rhadoop processing and analysis of data resources to the upper other business applications provide data analysis and knowledge discovery.

The lowest layer of platform is big data analysis data resources, the top layer of platform is user's needs, integrate R and Hadoop build an data analysis is the core processor. This architecture has many advantages, first, through Hadoop data manage system deal with big dataset, overcome R language statistical and analysis of large datasets memory bottleneck[13]. Second, make full use of R language data visualization and efficient model validation capabilities, sufficient to cover the Hadoop data analysis. Last, through R language data analysis packages provide functions,the platform can have many advance deep-analytical capabilities.

C. Data Processing Mechanism of Platform

The platform architecture can decompose big data analysis task ,many big data analysis can decompose into "big data part" and "small data analysis part".this idea is very efficient,

"big data part" completed by the distributed Hadoop MapReduce and "small data part" complete depth processing and analysis in R[14]. As the figure 2 platform architecture shows that R is above the distributed storage Hadoop ,there is a R bridge data access compont in the middle of R and Hadoop. R as platform data exchange processing environment,data resources storage in the Hadoop Distributed FileSystem(HDFS) Rhadoop component not only provides high-performance query processing Hadoop data mechanism but also allow perform R command to access data, then receive the data result from the distributed cluster environment. At the same time allow the Hadoop work node to run Rprogram for data analysis[15].Data can free conversion between R and Hadoop two platform, complete data analysis task decompose two part, data processing flow like the figure 3 shows in the following.

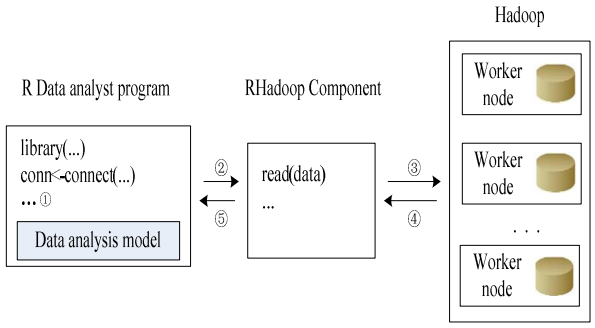


Figure 3. Data Analysis Process of Platform

IV. PLATFORM APPLICATIONS

DL abundant original big data resources, extremely have the value of analysis, featured resource database is a reflection of data utilization. In this section,we will take the DL county economy original dynamic datasets into consideration. An example to describe the details of the three main components of the platform specific analysis process.

A. Time Sequence Analysis

Dynamic dataset is one type evolving data that manage by the main line of time dimension ,this type data reflect the characteristic that data feature change with time[16]. Time sequence analysis is one statistical method for analyzing dynamic data processing in R language, It can reflect a set of data variation with time through data visualization[17]. consider the following table-1 simple example.

Table-1 Simple Data Example

	year	value1	value2
1	2011	212238	3205928
2	2012	440735	2109172
3	2013	616766	2514278
4	2014	213481	2929581

(Value1 represents Cumulative annual public budget revenue, unit ten thousand.
value2 represents Gross annual regional production, unit Millions)
The following is the R program to solve above problem, like this:

```
DF<-data.frame(year,value1,value2)
y <-zoo(DF)
y1 <-zoo(DF$value1,DF$year)
y2 <-zoo(DF$value2,DF$year)
plot(y,xlab ="year",ylab="budget revenue")
plot(y,,xlab ="year",ylab="Gross product")
```

Simple dynamic analysis of economic data, the annual budget and total output value data read into data frame, use R time sequence analysis data dynamically changing, as the figure 4 shows sequence data changes result.

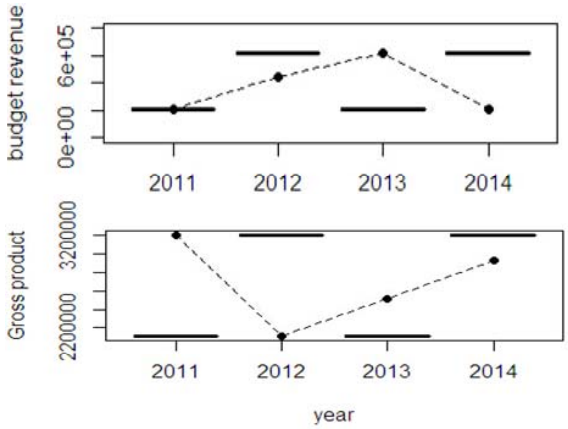


Figure 4. Time Sequence Data Changes

In this section, we introduce a simple example of time sequence analysis, through this dynamic data analysis with R; we can know that data is loaded into the R environment, then use R language to their strengths will be the visual representation of data, reflect the time main change dynamic data.

B. Application of Large Data sets

DL county economy dynamic data resources has many indicators, data comprehensive and complex, time-sensitive, include minimal assurance, total output value , budget, etc. About dozens of statistical indicators. For this type data resources analysis and use contribute to economic forecasting analysis risk control and decision-making guidance. In this chapter we will take this in practice, give this example that about the internet to deal with large data repository analysis and processing advantages.

Country economy data adopt HDFS distributed data manage system to storage data, use MapReduce to solve the task of data processing and clustered distributed computing. In this situation, Hadoop distributed parallel computing to process large part of data and then summary of the results. R receive the result data from the Hadoop, complete the small part of the data to analysis in R, and the results show to the platform users[18]. For instance, dynamic economic data research the number of low, poverty guaranteeing payment amount with

total output value economies and budget revenue relationship, Predict the future situation. Involving variety indicators of data, compared with the previous the platform to significantly boost performance. R sends the requested data to the cluster for summary data dimensions sum and then returns to R statistical model for further analysis. According to the county's economic have been observed data choose the fitting model and use OLS regression formula(1).

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_K X_{ki} \quad i = 1 \cdots n \qquad (1)$$

Use above formula fitting the data model to predict trends, in this section we use polynomial regression in R to improve the prediction accuracy regression.

```
fit<-lm(product ~low + I(low^2),data = df)

summary(fit)

plot(df$low,df$product,lty.smooth=2,pch =19,xlab="Guaranteeing payment",ylab = "Gross Regional Product" ,main ="Regression analysis")

lines(df$low,fitted(fit),col="red")
```

For the indicators data extract from the big dataset ,can we use lm(formula,data) multiple regression model, then show the results of the model. The data results shows that Guaranteeing payment of gold and total output value have the following (2) prediction formula, as the figure 5 shows data polynomial regression curve.

$$product = 1.322*10^{6} + 6.125*10^{2} low - 4.886*10^{-2} low^{2} \qquad (2)$$

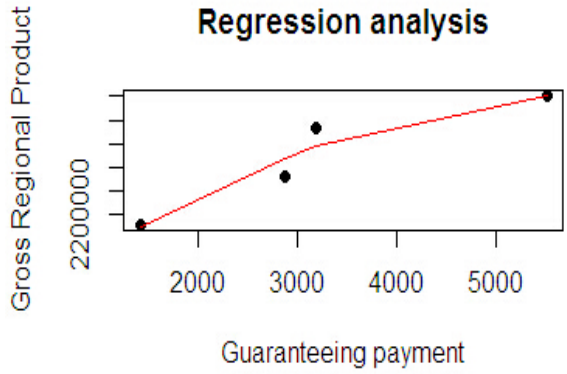


Figure 5. Regression Curve Between Low and Regional GDP

In this section, from the above analysis, mainly based on the DL county economy dynamic data. As an example to introduce R run in distributed big data analytics platform, we take this to verification platform whether it complete the expect goals when it build.

V. CONCLUSION

Now, under the DL open access mode, big data resources has needs of platform analysis .through use R and Hadoop to combine can built a scalable platform with deep-analytic, in this architecture , it can provide us statistical analysis environment to solve the difficult problem analyzed using original big data. In third section, we introduce the platform architecture and combine R with Hadoop Big data processing mechanism, in the last, we give a example to explain more about the R data processing platform in detail. Large complex data calculation process is completely transparent to the user in all the application platform, make user far from the tedious model calculations, improve resource utilization of the DL.

Acknowledgements

Fund project: Department of science, GuiZhou Province, Soft science fund projects (Qian Ke He Ti R Grant No. [2013]LKC2010, [2012] LKC2012)

Guizhou University of Finance and Economics university library characteristic database construction project

References

[1] Chuanful chen , digital library in the big data age, the 2012 annual meeting of the library of China.

[2] Sudipto Das,Yannis Sismanis,Kevin S.Beyer,Rainer Gemulla, Peter J.Haas,John McPherson .Ricardo:Integrating R and Hadoop,pages 987-998.2012

[3] Ron Daniel, Carl Lagoze, Sandra D.Payette. A Metadata Architecture for Digital Libraries.1998

[4] James Manyika, Michael Chui, Diana Farrell, Steve Van Kuiken, Peter Groves, Open data: Unlocking innovation and performance with liquid information.2013.

[5] The flow of information. http://blog.sina.com.cn/doctorwu11. 2015

[6] Anthes G.ACM Launches New Digital Library, J. Communication of the ACM,2011(2):23-24

[7] Chuanfu chen, ou qian,yuzhu dai. Study on the Construction of Digital Library in the Age of BigData, J. Books intelligence work.2014,58(7): 41-43

[8] TheFourthParadigm-Data IntensiveScientificDiscovery,Microsoft, 2009

[9] Anil Kumar Dhiman. Knowledge Discovery in Databases and Libraries.DESIDOC Joumal of Library&Information Technology pages 446-451 .2011

[10] Ru jiang bai,fu hai Leng.The research of integrate scientific data in big data. In library and Information Science .pages 94-99.2014

[11] Jin Yin Chu Juan.The Innovation and Development of Library Service—"Smart Analysis" in Big Data Era. pages 44- 46,37. 2013

[12] Giberti, Bruno. "The Classified Landscape: Consumption, Commodity Order, and the 1876 Centennial Exhibition at Philadelphia." Ph.D. Dissertation, University of California, Berkeley, 1994.

[13] R website,.http://www.r-project.org/. 2015..

[14] Dr. John McPherson, Ph.D.Data Intensive Analytics with Hadoop: A Look Inside,2010.

[15] Vignesh Prajapati. Big Data Analytics with R and Hadoop Pages 112-117. ,2013

[16] Robert Kabacoff. R in action Data Analusis and Graphice with R .M. Manning Publications 2013, pages 158-166.

[17] Time series of R language. http://doc.datapanda.net/a-Little-Book-of-R-for-Time-Series.pdf .2015.

[18] N.F.Samatova.pR:Introduction to Parallel R for Statistical Computing. In CScADS Scientific Data and Analytics for Petascale Computing Workshop, pages 505-509,2009.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Data Analysis Algorithm of Missing Point Association Rules for Air Target

Jiang Surong
Fourth department
Air Force Early Warning Academy
Wuhan, China
E-mail: yyhandy@yeah.net

Lan Jiangqiao
Fourth department
Air Force Early Warning Academy
Wuhan, China
E-mail: sqqking@163.com

Yang Yuhai
Fourth department
Air Force Early Warning Academy
Wuhan, China
E-mail: yyhandy@yeah.net

Abstract—It is important to analyze missing point phenomenon in early warning. By using data mining method, the association rules between air target missing point and status of early warning equipment can be concluded. A new mining algorithm is proposed, which firstly divided the target track into two categories, and then acquired the target air track net units with the same characters by clustering. Through matrix calculating and filtering false correlation sets, the association rules can be found. Experimental results demonstrated that this algorithm is efficient and accurate to mine the association rules among missing point events.

Keywords- early warning; data analysis; data mining; confidence

I. INTRODUCTION

Apriori algorithm can generate a large number of candidate item sets, and it needs frequent access to the database. In the literature [1], a new algorithm is proposed to find the minimum time interval of maximum correlation, which can reduce the complexity of the subsequent data mining by using the maximal group to divide the massive data. In the literature [2], multi class FP-tree is used to decompose the multi class rules, and the classification is predicted by the combination of multiple association rules. The computational complexity is reduced. The paper puts forward the improvement of the framework of multi layer association rule mining, and the ability of discovering association rules in abstract layer is stronger in the literature [3]. There are also researchers using the mining algorithm to mine the data of the alarm data, obtained a good auxiliary decision-making basis in the literature [4]. In wireless sensor networks, literature [5] uses association rule mining algorithm to discover the useful association between nodes and eliminate the redundancy of information between nodes. Literature [6] uses prediction and pruning strategy to reduce the number of conditional FP Tree in support of relatively small case that the algorithm can achieve higher efficiency. In the literature [7], association rules based on graph mining algorithm constructs an association graph based on, the candidate frequent item sets to construct an adjacency matrix, and then verify to determine whether the frequent item sets. In the literature [8], the author studies and improves the algorithm of association rule mining based on matrix, and reduces the data volume by layer by layer scanning. In the literature [9], in order to reduce disk I/O operation, the author has improved the computation method of the frequent item sets by divide the large database in advance. In the literature [10], the author puts forward a new method for the association rules mining, which is used in the time as a constraint, and has many applications in the economy and the weather. In literature [11], the author proposes to found the same characteristic of network element alarm group by clustering in telecommunication network database, then statistic based on correlation.

The above algorithms also has some shortcomings when applying them in the missing point database, because these algorithms are under the condition of the high support , highly reliability to mining association rules. In order to solve the above problem, we present a data mining algorithm with high reliability and high correlation degree. The algorithm can find association rules for those air tracks missing points that occur frequently; It can also find association rules for those air tracks missing points that do not occur frequently. Thus, the completeness and accuracy of the association rules for missing points can be improved.

II. AIR TRACK MISSING POINT ASSOCIATION RULE MINING ALGORITHM

The algorithm is based on two hypotheses.

Hypothesis 1: in general aerial, an air track missing point event reflects the characteristics and performance of the early warning radar equipment and characteristics of target. The same early warning radar and the same type of air targets will have the similar missing point rules. Thus, the correlation is similar when similar air target has similar missing point of air track.

Hypothesis 2: when multiple air targets execute one task, a cooperative relationship is formed. The flight state of these targets and the actions they took has some relationship. The closer the distance between them is, the higher the correlation degree is. Otherwise, the correlation degree is lower.

The basic idea of the algorithm for mining association rules is to find a similar missing point group of the air track by clustering. Finding a similar missing point group can predict the relationship between each missing point type.

2.1 missing point rule mining algorithm input

Missing point rule mining algorithm has two types of inputs. One is the independant air track has missing points, its type is A_1. The other is air tracks have missing points which have a group relations, its type is A_2.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.82

2.2 missing point rule mining algorithm output

Missing point rule mining algorithm's output is association rules set S which is related with the type of missing point of air tracks. Association rules in set S accordance with the relevant degree from high to low order.

2.3 core function module of missing point rule mining algorithm

(1) Cluster and grouping

Through cluster grouping, it can generate L class air track group, including N types of missing point.

For an independent track, it can find similar air tracks of the same type by clustering, such as a target of the same type, or the same airborne radar relative orientations, or the same aircraft radial velocity.

For cooperating tracks, such as a group of fighters, fighter emission missile, a regional aircraft in red and blue against, it can do relation clustering mainly through the same distance relation, or the same height relation, in order to generate a target air track group.

(2) The relevant set of missing points

After different target air track groups are generated, the phenomenon of missing point in air track groups can be mined. When the correlation is calculated, the missing point phenomenon related to a high degree of missing points set $C1$ can be generated.

Similarly, mining the rest of the missing point phenomenon in the group forms missing points set $C2$, $C3$,..., CL.

On the missing points set $C1$, $C2$, $C3$,...,CL, it can be do a collection addition operation. Thus, the relevant set C of missing point can be calculated.

Finally filtering and rules determining can be applied on the relevant set C of missing point respectively. Then the rule set S of missing point can be got.

III. STUDY ON THE KEY PROBLEM OF MINING ASSOCIATION RULES FOR MISSING POINT

In practical application, only the information closely related to the missing point phenomenon is selected as a data mining keyword in the missing point database, such as missing point occurrence time, and the serial number of the air track, and missing point type description.

Definition 1: Missing point time window

The window $W=(w, t_s, t_e)$ refers to one missing point sequence in sets $S=(s, T_s, T_e)$.

Here, $s=\langle(A_1, t_1), (A_2, t_2), \ldots, (A_n, t_n)\rangle$ is couples composed of multiple orderly missing point events.

If E is the given missing point type set, then $A_i \in E$, t_i is the missing point occurrence time. T_s and T_e represent the initiation time and termination time of missing point of air track. Window W is composed of all the missing point (A, t) which meet $T_s \leqslant t \leqslant T_e$ in missing point sequence S.

The limiting condition is $t_s \langle T_e$ and $t_e\rangle T_s$.

The width of the missing point time window is $w=T_e-T_s$.

The choice of window width has a certain effect on the time granularity of association rules, so it must be determined carefully. The problem of determining the time window width is equivalent to determining the maximum time interval of the related event in the application.

The sliding window size is also an important parameter. In the application, it refers to missing point event interval. When the database is traversed, the greater the sliding step size is, the higher efficiency of the mining algorithm has, but the accuracy of the rules will be reduced. If the sliding step size becomes smaller, the efficiency of mining algorithm will be reduced, and the cost of the system will be increased, but the accuracy of association rules will be improved.

In order to make the adjacent windows have certain overlap, the window sliding step size is generally less than 1/2 of the window size.

Definition 2: The missing point matrix

Assume that there is missing point vector R_i (i=1, 2, …, N), N is the total number of missing point types. R_i is composed of n (n≤N) missing point row vectors composed of ordered missing point events.

The matrix $D_i(m*n)$ is the missing point matrix of missing point vector R_i.

$$\mathbf{D_i} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{12} & d_{22} & & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \tag{1}$$

The rows present missing points in all sliding windows. The columns present various missing point events. Matrix element is the occur frequency of each missing point. For example, D_{mn} presents the occur frequency of the missing point event A_n occurring in the m window.

Missing point matrix D_i is generated by m missing point vectors. Through calculating correlation to mine missing point events related with A.

The missing point event has two parts in missing point database. They are missing point air track identification and missing point type description. Missing point air track identification includes air track sequence number, target attribute, target number, etc. The missing point type description includes missing point type, missing point causes, severity, etc.

Definition 3: Correlation degree

Correlation degree is the correlation coefficient, which is mainly used to describe the distribution of two random variables.

$$\operatorname{cor}(A_i, A_j) = \frac{\operatorname{cov}(\mathbf{R_i}, \mathbf{R_j})}{\sigma_i \sigma_j} = \frac{\sum_{i,j=1}^{m} (w_i - \overline{w_i})(w_j - \overline{w_j})}{\sqrt{\left(\sum_{i=1}^{m} (w_i - \overline{w_i})^2\right)\left(\sum_{j=1}^{m} (w_j - \overline{w_j})^2\right)}} \tag{2}$$

Among them, the weight value w_i, w_j refers to the frequency of missing point vectors R_i, $R_j(i, j=1, 2,...$ N). N is the total number of missing point type. m is the

total number of sliding window and $\overline{w_i}$, $\overline{w_j}$ is frequency mean of the missing point vector R_i, R_j.

Definition 4: Correlation between the missing point events

According to the experience, the farther the geographical distance between air targets becomes, the smaller probability there is between the missing point tracks. According to the attribute information of the air track, the distance between the 2 targets can be calculated.

The correlation degree between the missing point events is defined as:

$$c(A_i,A_j)=\mathrm{cor}(A_i,A_j)/\sqrt{w_{i,j}} \qquad (3)$$

Among them, the $w_{i,j}$ represent the weight value of geographical distance between A_i and A_j .

Definition 4 and definition 3 can deduce

$$c=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{12} & c_{22} & & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \qquad (4)$$

The row and column in two dimensional correlation matrix c（n*n） are both missing point types. The matrix elements c_{ij} is the normalized correlation degree calculated from the weight value $w_{i,j}$ and w_i .

A. Correlation set of multiple missing point vectors

There are two kinds of methods for calculating correlation set of multiple missing point vectors.

(1) Every 2 vectors high degree of vector will be obtained as a whole, and then the correlation degree with other vectors is calculated. In this method, when the number of two sets is large, the computational complexity is relatively large.

(2) To combine of high degree of correlation.

The algorithm complex degree of calculating vector correlated matrix coefficient is $O[mn(n\text{-}1)]$. In the practical application, m is much larger than n. In order to improve the efficiency of the algorithm, the second method is used to produce missing point vector of multiple items.

B. False correlation set filtering method

Lemma 1: All non empty subsets of correlated sets are correlated.

It can be proved by contradiction. Assuming A_i, B_i, C_i are missing point events, c_{min} is the minimum correlation threshold and missing point vector is $I=A_iB_i$.

If $c(I)=c(A_iB_i)<c_{min}$, then $I=A_iB_i$ is not relevant, that is, A_i and B_i are not related.

Therefore the missing point vector $I'=I\cup C_i$ must be smaller than the threshold c_{min}. That is$c(I')=c(I\cup C_i)<c_{min}$. So $I'=I\cup C_i$ is not relevant.

Corollary: if A_iB_i, A_iC_i and B_iC_i are related, it is considered that the $A_iB_iC_i$ is relevant.

However, according to the relevant experience, in the field of air defense warning, this corollary is not tenable. So some false correlation sets with low correlation are produced often.

The false correlation set can be filtered by rule determination method. The specific method is to merge some similar degree of correlation first, and then filter.

Filtering rules are described as: $c(A_iB_i)=X_1$, $c(A_iB_iC_i)=X_2$. For a given correlation threshold c_{min}, and an given arbitrarily small positive number ε, if $X_1>c_{min}$, $X_2>c_{min}$, and $|X_1\text{-}X_2|<\varepsilon$, after joining the missing point event C_i, does not have an impact on correlation, that is to say, correlation set $A_iB_iC_i$ is not provide more information. Hence the correlation set can be filtered.

Definition 5: Credibility

Credibility defined as the conditional probabilities that missing point event B also occurs when missing point event A occurs within the given time window. That is:

$$\mathrm{conf}(A\rightarrow B)=|A\cap B|/A=P \qquad (5)$$

In the form of rules: after missing point event A occurs, the time window for the w interval, the probability of missing point event B occurs is P. The credibility can give a probability based reliability characterization of the association rules.

IV. EXPERIMENT AND ANALYSIS

In the practical application, we have carried out the experiment and make contrast analysis of this algorithm[12]. The experimental data is missing point database (200000 records data) produced by one year flight and detected. In the experiments, we take missing point events (by the serial number of air track, and the corresponding air target type, the true/false attribute of the air track, type of missing point, missing point level) and missing point occurrence time (in seconds) as a keyword for data mining. Missing point time window is set to 5 minutes, the moving step is set to 1 minute. Minimum support is divided into 5 levels, respectively, 1%, 2%, 3%, 4%, 5%.

In the case of different minimum support, the MPAR algorithm and the WINEPI algorithm are compared. The results are shown in Figure 2. The comparison results in the correlation sets / frequent sets are shown in Figure 3.

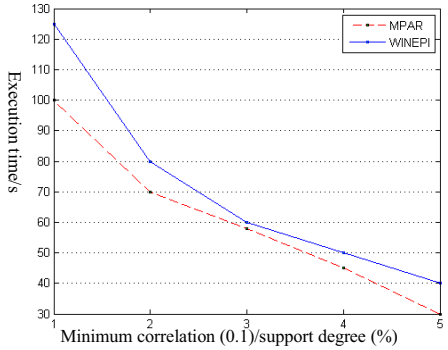


Figure 2 comparison results of the execution time of the algorithms

As can be seen from figure 2, the execution time of the 2 algorithms decreases with the increase of the minimum correlation degree and support degree, because the WINEPI algorithm candidate frequent set

length increases every time to scan the database, so the time overhead is relatively large. Based on the statistical correlation algorithm, it only needs to scan the database once, then the matrix calculation can be carried out, so the execution time is less.

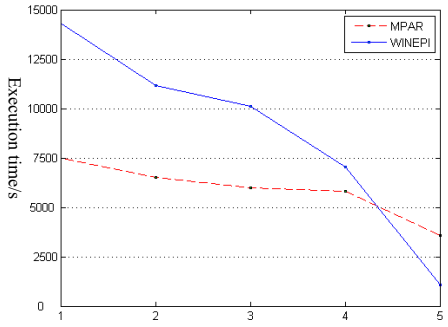


Figure 3 rule set comparison results

From Figure 3 it can be seen that with the increase of the minimum correlation, the number of the eliminated sets increases. WINEPI algorithm decreases rapidly with the increase of the minimum support degree. But the correlation algorithm based on statistical decreases slowly with the increase of the minimum correlation degree, because when the WINEPI algorithm considers the execution efficiency, first of all, it filters out according to the minimum support a large number of non frequent missing point event and can't find infrequent but correlated missing point rules. For example, some cases can't be detected when the degree of occurrence is low, but the number of missing point is large. But based on statistical correlation algorithm for missing point event first classifies, then establishes the association matrix, and finally, mine the association rules according to correlation between the missing point events. This avoids the impact by the missing point frequency. In the case of the credibility conf > 0.8, the mining rule set is shown in table 1 and table 2.

TABLE I PARTIAL ASSOCIATION RULE (WINEPI ALGORITHM)

Rule	Missing point pair1	Missing point pair2	Support
1	1001,L1	1010,L1	0.58
2	1223,L1	1224,L1	0.47
3	1335,L2	1468,L2	0.47
4	1778,L3	1993,L3	0.43
…	…	…	…

In Table 2 of the rules in 3 cases, we can find the meanings as follows: the probability of missing point pair1 (1335,L2) and missing point pair2 (1468,L2) emergence in the same 5 minute window is more than 98%, and the credibility of this rule is more than 80%.

It can be seen from the rule result set that WINEPI algorithm is limit with a minimum support degree. Only mining association rules between frequent missing points can be found with long track correlation missing point relationship, as shown in Table 1. Using algorithm based on statistical can not only find out the long track between missing point association rules, but also can find that short track between missing point association rules, and long track and short track between missing point related rules, as shown in Table 2.

TABLE II PARTIAL ASSOCIATION RULES (MPAR ALGORITHM)

Rule	Missing point pair1	Missing point pair2	Support
1	1001,L1	1010,L1	0.8
2	1223,L1	1224,L1	0.7
3	1335,L2	1468,L2	0.6
4	1440,L3	1441,L3	0.5
5	1448,L3	1451,L3	0.5
…	…	…	…

V. CONCLUSIONS

The paper researched association rules mining for air track missing point phenomenon in the early warning detection field. It presents a statistical based correlation missing point association rule mining algorithm (MPAR). The algorithm to the high degree of correlation, highly reliability condition, track for independent track and group track were mining, can also dig out the association rules between frequent and non frequent leak sequence, improve the early warning and detection accuracy and wholeness of the leak data analysis.

REFERENCES

[1] Wang Ning, Yang Yang, Gong Huarong. Mining Method of Key Time Interval Based on M aximum Clique[J]. Computer Science, 2012, 39(6): 166-169.

[2] Li Bo, Li Hong, Wu Min. Multi-label classification algorithm based on association rules[J]. Control and Decision, 2009, 24(4): 574-582.

[3] Cai Hongguo, Peng Yizhong. multiple-layers association rule mining algorithm based on GEP and its application[J]. Computer Engineering and Design, 2010, 31(1): 137-140.

[4] Zhou Guangning, Xu Yaguo. Data processing based on association rules analysis of thepolice[J].Police Technology, 2010, 5:13-15.

[5] Lin Yaping, Mei Shuying, zhou Siwang. A Lexicographic Tree Algorithm forMining Association Rulesfrom Wireless Sensor Networks[J]. Computer Engineering and Science, 2010, 32(4): 119-124.

[6] Qian Xuezhong, Hui Liang. Mining algorithm of maximal frequent pattern based on improved FP-tree in association rules[J]. computer Engineering and Design, 2010, 31(21): 4635-4638.

[7] Xu Benyu, Wang Yanbo. A Mining Algorithm for Association Rules Based on Graph and Matrix[J]. Journal of Chongqing Communication Institute, 2010, 29(3): 53-55.

[8] Wang Juanqin, Li Shuqin. Research and Improvement on an Algorithm of MiningAssociation Rule Based on Matrix[J]. Computer Measurement and Control, 2011, 19(9): 2275-2277.

[9] Cui Jian, Li Qiang, Wang Guoshi. Algorithm of Association Rule Mining for Large Transaction Databases[J]. Journal of Air Force Radar Academy, 2011, 25(3): 205-208.

[10] Liu Yinghua, Li Guangyuan, Liu Yongbin. An Efficient Mining Algorithm ofTemporal Association Rules[J]. Computer Engineering and Science, 2011, 33(9): 105-108.

[11] Xu Qianfang, Xiao Bo, Guo Jun. An alert association rule mining algorithm based on correlation statistics [J]. Journal of Beijing University of Posts and Telecommunications, 2007, 30 (1): 66-70.

[12] hatonen K.Knowledge discovery from telecommunication network alarm databases[C].ICDE'96. New Orleans,1996:115-1

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Comprehensive Evaluation System of Association Rules Based on Multi-index

Shunli Ding, Xin He, Hong Liang
Dept. of Information Engineering
Environmental Management College of China
Qinhuangdao, China
dingsl@163.com, hex80@163.com, 40968853@qq.com

***Abstract*—Data mining is to extract the potentially useful knowledge and information from large amounts of data. How to dig up effective, reliable, understandable, and interesting association rules from vast amounts of information to help people make decisions has become an urgent problem to be solved. People want to use a reasonable evaluation method to measure reliability or validity of association rules, producing more interesting rules for the users. A multi-index comprehensive evaluation system of association rules was constructed, which can evaluate the association rules from multi-angle and multi-dimensional. Giving different weights for each index, more valuable association rules can be found. Practical data make it clear that the comprehensive evaluation system is rational and superior.**

Keywords-data mining; association Rule; Comprehensive Evaluation; Multi-index

I. INTRODUCTION

Association rules analysis is a widely used method in data mining and one of the common forms of knowledge about the relationship between different things, which is easy to be understood and accepted. So far, most of association rules mining method are based on support - confidence - lift framework. However, in the real world, data is ever-changing. It is found that to choose suitable minimum support and minimum confidence threshold is not easy, when digging for some data sets. Therefore, many scholars begin to pay close attention to the evaluation of association rules [1-3]. People want to use a reasonable evaluation method to measure reliability or validity of association rules, producing more interesting rules for the users. However, the conventional evaluation methods evaluate the rules from a point of view which have some limitations. Besides, the results of the various evaluation methods have certain deviation, so it is difficult for decision makers to make decision when faced with a choice. In order to solve these problems, this paper puts forward a comprehensive evaluation system based on multi-index.

II. ANALYSIS OF VARIOUS EVALUATION INDEXES OF ASSOCIATION RULES

At present, so much association rules mining algorithms frequently generate a large number of rules, but only a small number of the rules may be selected for implementation by decision makers due to the limitations of resources. Therefore, the evaluation of association rules has become an important research topic in the field of study. The evaluation index about if an association rule is valuable mainly consist of the support, confidence, and lift for the comprehensive evaluation system.

(1) Support. Support is the percentage of transactions in total transactions that contain antecedent and consequent of rules. In general, the higher support degree, the higher statistical significance of a rule, that is, the rule will be widely used. However, support as a measure index has a shortcoming that is subject to the sparsity of item. As a result of high threshold of the support degree, some of the potential and valuable rules may be ignored.

(2) Confidence. Confidence is taken to be the conditional probability that the consequent occurs under the conditions of the antecedent. The greater confidence degree is, the higher reliability of the rule is. However, the drawback of confidence degree is that policymakers tend to be misled when faced with negative rules.

(3) Lift. Recent researches and experiments demonstrate that using the traditional evaluation system to assess rules has a lot of limitations. For example, they may generate a great many trivial association rules, many of which are not interested for the users, and may even contain redundant, reduplicative and meaningless. Therefore, to mine more valuable association rules, interest is the most widely used and the highest recognition criteria. In [4], interest is to be known as lift. Lift refers to the probability ratio between support and the emergency of consequent of rules. The lift reflects the relevant relationship between the antecedent of rules and the consequent.

III. A STRUCTURE OF WEIGHTED COMPREHENSIVE EVALUATION SYSTEM

Each evaluation index is a reflection of a certain perspective on association rules, if using comprehensive evaluation method combines a variety of evaluation indicators, to evaluate association rules from various angles, more effective rules may be dug out. Therefore, an evaluation structure of association rules is established by using the weight. First, use association rules mining algorithm such as Apriori algorithm for data association rule mining, then use the appropriate method to give the corresponding weights of evaluation indexes, finally use the weighted comprehensive evaluation model to get association rules, so as to get the final total sorts of association rules. The framework of the comprehensive evaluation system is shown in Figure 1.

The determination of weights is a complicated problem. According to the different areas have different weights, for example, if the application domain is business domain in

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.83

which it is concerned with the application of rules, then the weight of support degree is relatively large, but if it is in the field of science area, the reliability required is relatively high, therefore, the weight of confidence is large, but the other is small. And there are many methods to determine the weights such as Delphi Method, analytic hierarchy process (AHP), and neural network method, etc.

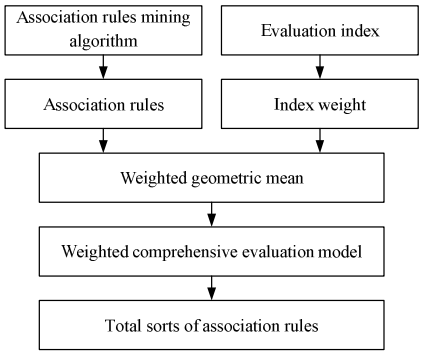


Figure 1. The framework of comprehensive evaluation system

Comprehensive assessment of each index also has a variety of methods, such as the simple weighted sum method, principal component analysis, weighted quadrature, and weighted geometric average method.

The total ranking of association rules can be calculated according to the model, getting a unique comprehensive evaluation result, which may facilitate the decision makers choose more valuable rules for application.

IV. THE TESTING AND ANALYSIS OF COMPREHENSIVE EVALUATION SYSTEM

A. *Experiment data*

The experiment data is from domestic professional research data service hall. Datatang [5] is committed to provide higher education institutions home and abroad, scientific research institutions, research companies and related researchers with data and added-value services on the direction of research or application requirements. All the data are with high reliability and authenticity. The data have recorded 1000 customer purchasing records, containing 20 items like hering, olives, ham, turkey, bourbon, ice_crea, and so on.

B. *Experiment process*

In the following experiment, first of all, analyze association rules in the experimental data by use of SPSS clementine, and carry on the concrete explanation of the mined association rules. Secondly, through the mining results, prove the rationality of comprehensive evaluation based on correlation analyses. Get the comprehensive evaluation results by calculating again, then analyze and compare the comprehensive evaluation results with those evaluation results of single one, illustrating the superiority of comprehensive evaluation system.

C. *The single index evaluation*

With only one index to evaluate the association rules, the evaluation results of each method are different, and the gap of value is insignificant. That is to say, the results obtained by a single evaluation index do not have a significant correlation. In order to facilitate the analysis and research in mining, the largest number of the antecedent and consequent of rules set as 1, which is only mining association between the two pieces of goods. Through repeated experiments, the mining results when minimum support degree is 10% and minimum confidence is 60% are selected for research. Digging out 20 association rules, the specific situation of each index are shown in table .

Through the analysis, evaluating of association rules with only a single evaluation index has some shortage as followings.

(1) The single evaluation indicator evaluates rules from an angle. For example, support can only represent the probability of rules appearing, and confidence that reliability of rules, the lift reflects the relevant relationship between the antecedent of rules and the consequent. These three indicators are only from one side to evaluate rules.

(2) According to the different evaluation methods, the results of each single evaluation index are different, as shown in figure 2. But in practice, due to the limitations of resources, only a small number of the rules may be selected for implementation by decision makers. When facing different evaluation results, the decision maker has some difficult to make a choice. So it is necessary to carry out a comprehensive evaluation on the rules, so that the only evaluation result can be obtained, helping decision makers to choose the most valuable rules.

(3) The results obtained by single index evaluation are not significantly different, in other words, is not significant between the valuable rules and useless rule. For example, in Figure 2, in accordance with the degree of support in descending order, rule *T1* is the most value rules, and the rules of *T20* is the most worthless rules. However, if in accordance with the confidence or lift ranking, the rule *T20* is to achieve high ranking, meaning that *T20* is of some value. Therefore, only using a single evaluation index is not a really good way to judge value of rules.

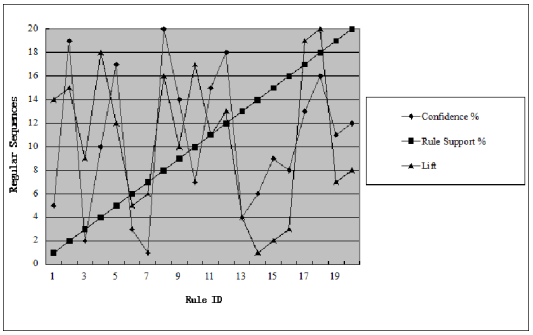


Figure 2. The comparison of single index evaluation results

D. Comprehensive index evaluation

The comprehensive evaluation system based on index weight can be determined using a variety of methods, and here uses AHP as an example to illustrate. First of all, construct judgment matrix. This paper adopts the "support - confidence - increase degree" making comparison between any two indexes, and construct judgment matrix between two indexes.

Secondly, calculate the relative important degree of every index. Here using square root method to calculate the important degree of each index. A new vector is obtained by multiplying the rows of matrix A, then to get the final matrix based on normalization of the new vector whose components are divided to the power three. The weight of each index can be calculated.

TABLE I. THE MINING RESULTS OF SINGLE INDEX ASSOCIATION RULES

ID	Consequent	Antecedent	Confidence %	Rule Support %	Lift
T1	heineken	cracker	74.486	36.164	1.247
T2	cracker	heineken	60.535	36.164	1.247
T3	heineken	soda	79.938	25.874	1.338
T4	heineken	baguette	65.903	25.874	1.103
T5	olives	bourbon	60.628	25.075	1.272
T6	cracker	soda	77.16	24.975	1.589
T7	heineken	artichok	80.844	24.875	1.353
T8	cracker	bourbon	60.145	24.875	1.239
T9	hering	baguette	62.85	24.675	1.295
T10	heineken	avocado	67.68	24.476	1.133
T11	hering	corned_b	62.404	24.376	1.285
T12	olives	corned_b	60.614	23.676	1.272
T13	olives	turkey	76.389	21.978	1.603
T14	ice_crea	coke	73.81	21.678	2.287
T15	coke	ice_crea	67.183	21.678	2.287
T16	avocado	artichok	67.208	20.679	1.858
T17	heineken	chicken	64.038	20.28	1.072
T18	heineken	sardines	61.433	17.982	1.028
T19	hering	steak	65.652	15.085	1.352
T20	olives	steak	64.348	14.785	1.35

Thirdly, consistency checking. Consistency checking is implemented. After examination, the consistency ratio C.R. is far less than 0.1, and satisfies the requirement of consistency, so the weight obtained is effective.

Obtained by AHP method, C, S, L weights of three indexes are 0.110, 0.582, and 0.308. The weight of confidence is largest, which is considered to be the most important index for evaluating association rules. As a result, the comprehensive evaluation coefficient obtained is shown in formula (1).

$$R=C^{0.11} \times S^{0.582} \times L^{0.308} \quad (1)$$

In order to illustrate the superiority of the comprehensive evaluation system compared with the single evaluation indicators. For table data analysis, use the comprehensive evaluation based on weighted method to evaluate rules, using the formula (1) to calculate the results of comprehensive evaluation, as shown in Table Ⅱ.

The results can be seen from Table , weighted comprehensive evaluation method and the traditional evaluation methods have very big difference in the evaluation results. In addition, *T1, T2* is considered the most valuable by a single evaluation method and that *T17~T20* is of no value, and that the evaluation of these rules are consistent, but there are some differences in the evaluation results of other rules. It is metaphysical to evaluate rules with the single evaluation method merely from a perspective, therefore, a comprehensive evaluation method makes up for the deficiency of single evaluation method.

The comprehensive evaluation system has the following advantages.

(1) The original single evaluation methods are integrated into comprehensive evaluation system, which can evaluate the association rules from multi-angle, multi-dimensional, making the most use of evaluation results information. Combine many kinds of evaluation results, which make the evaluation results more reasonable. The comprehensive evaluation system can organic combination with various indicators, to evaluate the rule no longer isolated analysis.

(2) The Comprehensive evaluation system may amend the original single index based on geometric weighting method, making significant differences between valuable rules and useless rules. Therefore, only when the value of

each index association rules is very significant, the coefficient of assessment can reflect significant changes, which can eliminate the influence of incorrect manipulation.

(3) The comprehensive evaluation system is flexible. For different application areas, the importance of each index is different. Therefore, using the weighted comprehensive evaluation system can change the weight flexible decided by experts or users according to the situation.

TABLE II. **THE RESULTS OF COMPREHENSIVE EVALUATION**

ID	Consequent	Antecedent	R	comprehensive ordering
T1	heineken	cracker	57.326	1
T2	cracker	heineken	56.033	2
T14	ice_crea	coke	51.255	3
T15	coke	ice_crea	50.727	4
T6	cracker	soda	49.995	5
T3	heineken	soda	48.590	6
T7	heineken	artichok	47.714	7
T13	olives	turkey	46.483	8
T16	avocado	artichok	46.297	9
T5	olives	bourbon	45.569	10
T9	hering	baguette	45.566	11
T11	hering	corned_b	45.108	12
T8	cracker	bourbon	44.946	13
T4	heineken	baguette	44.823	14
T12	olives	corned_b	44.067	15
T10	heineken	avocado	43.882	16
T17	heineken	chicken	38.434	17
T18	heineken	sardines	35.219	18
T19	hering	steak	34.848	19
T20	olives	steak	34.352	20

(4) The comprehensive evaluation system is scalability. This paper just chooses three kinds of commonly used evaluation indicators, but based on this, it also can be easily integrated with other indicators.

(5) Evaluation system eventually presents a unified comprehensive coefficient to users, outputting the rules in accordance with the size of R, more convenient for decision makers to make comparison and choice.

V. CONCLUSION

Based on the study of the evaluation method, this paper puts forward a comprehensive evaluation system of association rules. The system can integrate a variety of evaluation results to evaluate association rules from multi-angle and multi-dimensional, eventually presenting a unified comprehensive coefficient of evaluation results to users. In the testing phase, the rules are evaluated by using three parameters with support, confidence and lift comprehensively, and comprehensive coefficients of each rule are calculated based on AHP. The experiment results have verified the rationality and superiority of the comprehensive evaluation system. In practical application, this method has flexibility and extensibility.

REFERENCES

[1] Elnaz Delpisheh, John Z. Zhang. “A Dynamic Composite Approach for Evaluating Association Rules”. 2011 Seventh International Conference on Natural Computation, 2011, pp.1893-1895.

[2] Benites F, Sapozhnikova E. “Evaluation of Hierarchical Interestingness Measures for Mining Pairwise Generalized Association Rules”. IEEE Transactions on Knowledge and Data Engineering, Volume:26 Issue: 12, pp. 3012 – 3025,2014

[3] Djenouri Y, Gheraibia Y, Mehdi M, Bendjoudi A. “An efficient measure for evaluating association rules”. 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 406-410, 2014

[4] S Brin，R Motwani，JD Ullman，S Tsur. “Dynamic itemset counting and implication rules for market basket Data”. ACM SIGMOD Record，26(2), pp. 255-264, 1997.

[5] http://www.datatang.com/datares/go.aspx?dataid =613168.

NIB2DPCA-based Feature Extraction Method for Color Image Recognition

Zongyue Feng
School of IoT Engineering
Jiangnan University
Wuxi, China
zongyue.feng@foxmail.com

Jiagang Zhu
School of IoT Engineering
Jiangnan University
Wuxi, China
zhujg@jiangnan.edu.cn

***Abstract*—A novel method for color image feature extraction is proposed. First of all, non-iterative bilateral projection based 2DPCA algorithm (NIB2DPCA) is employed to extracted feature information from three channels of a given color image respectively. Then the three extracted feature matrices are reconstructed to form a two-dimension middle matrix. After that, NIB2DPCA is employed again to extract features from the middle matrix to obtain the final features. Experimental results on CVL and FEI databases show that, contrast to existing similar methods, the recognition accuracy of our method increased by 7-16 percent.**

Keywords-component;color image recognition; eature extraction; Non-iteration bilateral projection based 2DPCA(NIB2DPCA); principal component analysis (PCA)

I. INTRODUCTION

In reality, image is almost colorful, and the color of image provides rich information for image recognition [1], so more and more researchers began to use the image color information to improve the accuracy of the recognition [2]. The experimental results of literature [3] show that, compared with gray feature based images, the color images contain more information for recognition, and that making full use of the color information can obviously improve the recognition accuracy of the color images.

The feature extraction is an important step in the process of image recognition, and PCA [4] is one of the feature extracting methods used frequently. After that, many new improvements on PCA were proposed [5]. In 2009, Guan proposed a Non-iteration bilateral projection based 2DPCA method (NIB2DPCA) [6] whose time of feature extraction is greatly shortened because both the left and the right multiplying projection matrices are calculated without iteration. NIB2DPCA is the fastest bilateral 2DPCA method at present.

In color image pre-processing, since a color image is generally represented in a three dimensional array in mathematics, we hope to transform the three dimensional color image array into a corresponding two dimensional matrix so that 2DPCA can be employed to extract features. At present, the methods for such transform mentioned above for recognition purpose are mainly as follows: the first method is to transform a three dimension color image array into a corresponding two dimension pseudo gray image matrix, according to the values of pixels in the same position of the three RGB layers, then, feature extraction and image recognition are performed in the same ways as traditionally used for a gray image [7]; the second method is to extract feature information from three RGB channels of a given color image respectively, then to classify or to recognize the image by the three feature matrices in some RGB-independent ways; the third method proposed by Wang in 2008 [11] firstly converts the three two dimensional matrices from three channels of a color image into corresponding three vectors, secondly merges the three vectors into a two dimensional matrix, thirdly extracts features using 2DPCA, and finally performs recognition using the nearest neighbor classification. The third method is the best one of the above methods from the viewpoint of utilizing color information. However, its memory space occupied is too much and so is its computation time.

Introducing NIB2DPCA in [6] into the color image recognition method in [11] and adding a feature pre-extraction step to it, we proposed a novel method for color image feature extraction in order to further improve the recognition rate. Experimental results on CVL and FEI color face database show that our method is obviously superior to the above methods.

II. NON-ITERATIVE BILATERAL PROJECTION BASED 2DPCA (NIB2DPCA)

Let M be the number of all training samples, and let $U \in R^{m} \times R^{l}$ and $V \in R^{n} \times R^{r}$ be the left and right multiplying projection matrix respectively, for $m \times n$ image matrix $A_i (1 \le i \le M)$ and $l \times r$ projected feature image $B_i (1 \le i \le M)$, the bilateral projection is formulated as follows:

$$B_i = U^T A_i V , \tag{1}$$

where left multiplying projection matrix is:

$$U = [U_1, U_2, ..., U_l] ,$$

and $U_1, U_2, ..., U_l$ are the l eigenvectors of C_U corresponding to the l largest eigenvalues respectively, C_U is formulated as follows:

$$C_U = \frac{1}{M} \sum_{i=1}^{M} A_i A_i^T . \tag{2}$$

Left multiplying projection matrix U only takes the column information of the images.

Right multiplying projection matrix is:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.84

$V = [V_1, V_2, ..., V_r]$.

Where, $V_1, V_2, ..., V_r$ are the r eigenvectors of C_V corresponding to the r largest eigenvalues respectively, C_V is formulated as follows:

$$C_V = \frac{1}{M}\sum_{i=1}^{M} A_i^T A_i \quad . \tag{3}$$

Right multiplying projection matrix V only takes the row information of the images.

After getting the left and right multiplying projection matrix respectively, we can obtain the feature matrix Y of any given probe sample image X by the bilateral projection. After that, some classification method can be employed through the comparison with similarity between Y and B_i.

III. NIB2DPCA-BASED FEATURE EXTRACTION METHOD FOR COLOR IMAGE

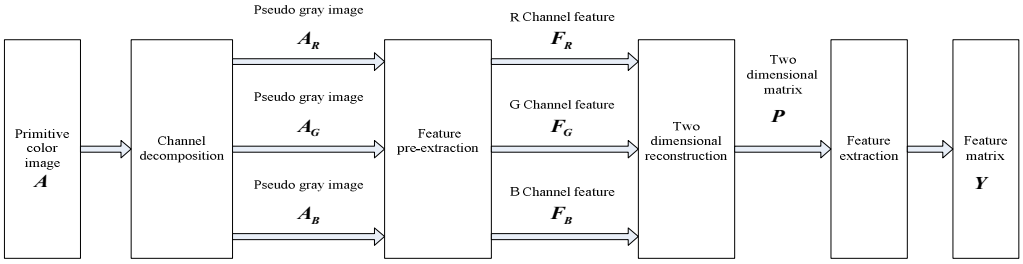


Figure 1. The proposed color image feature extraction method

Our novel feature extraction method is shown in Figure1, and the main process includes four steps: (1) channel decomposition, (2) feature pre-extraction, (3) two-dimension reconstruction, and (4) feature extraction.

For $m \times n$ color image A, the three R, G, B channels of A are respectively represented in matrix as follows:

$$A_R = \begin{bmatrix} a_{11}^{(R)} & a_{12}^{(R)} & \cdots & a_{1n}^{(R)} \\ a_{21}^{(R)} & a_{22}^{(R)} & \cdots & a_{2n}^{(R)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}^{(R)} & a_{m2}^{(R)} & \cdots & a_{mn}^{(R)} \end{bmatrix} \tag{4}$$

$$A_G = \begin{bmatrix} a_{11}^{(G)} & a_{12}^{(G)} & \cdots & a_{1n}^{(G)} \\ a_{21}^{(G)} & a_{22}^{(G)} & \cdots & a_{2n}^{(G)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}^{(G)} & a_{m2}^{(G)} & \cdots & a_{mn}^{(G)} \end{bmatrix} \tag{5}$$

$$A_B = \begin{bmatrix} a_{11}^{(B)} & a_{12}^{(B)} & \cdots & a_{1n}^{(B)} \\ a_{21}^{(B)} & a_{22}^{(B)} & \cdots & a_{2n}^{(B)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}^{(B)} & a_{m2}^{(B)} & \cdots & a_{mn}^{(B)} \end{bmatrix} \tag{6}$$

In channel decomposition step, let M be the number of all training samples, every training sample color image $A_i\ (1 \le i \le M)$ is respectively decomposed into three pseudo gray images $A_{R_i}\ (1 \le i \le M)$, $A_{G_i}\ (1 \le i \le M)$, and $A_{B_i}\ (1 \le i \le M)$ corresponding to the three channels.

In feature pre-extraction step, we want to obtain F_R, F_G, and F_B respectively from A_R, A_G, and A_B by using NIB2DPCA in section 2.

To obtain F_R from A_R firstly, viewing $A_{R_i}\ (1 \le i \le M)$ as a two dimensional image matrix respectively, we first calculate

$$C_{UR} = \frac{1}{M}\sum_{i=1}^{M} A_{R_i} A_{R_i}^T \tag{7}$$

and

$$C_{VR} = \frac{1}{M}\sum_{i=1}^{M} A_{R_i}^T A_{R_i} \quad . \tag{8}$$

Then we calculate

$U_R = [U_{R1}, U_{R2}, ..., U_{Rl}]$

and

$V_R = [V_{R1}, V_{R2}, ..., V_{Rr}]$

where $U_{R1}, ..., U_{Rl}$ are the l eigenvectors respectively corresponding to the l largest eigenvalues of C_{UR}, and $V_{R1}, .., V_{Rr}$ are the r eigenvectors respectively corresponding to the r largest eigenvalues of C_{VR}.

Finally, we get F_{R_i}:

$$F_{R_i} = U_R^T A_{R_i} V_R, \quad i = 1, \cdots, M \tag{9}$$

Similarly, we can also get F_{G_i}, F_{B_i} from A_{G_i}, A_{B_i} respectively.

In two-dimensional reconstruction step, we want to fuse $F_{R_i}, F_{G_i}, F_{B_i}$ to yield a two-dimension middle matrix $P_i\ (1 \le i \le M)$. That can be done easily using the following formula:

$$P_i = \begin{bmatrix} \left(vec\left(F_{R_i}\right)\right)^T \\ \left(vec\left(F_{G_i}\right)\right)^T \\ \left(vec\left(F_{B_i}\right)\right)^T \end{bmatrix} \quad , i = 1, \cdots, M, \tag{10}$$

where, $vec(\bullet)$ means vectorization.

In feature extraction step, NIB2DPCA is employed again to obtain final feature matrix Y_i from P_i, $i = 1, \cdots, M$. Again, viewing P_i as a two dimension image matrix, we first calculate

$$C_{UP} = \frac{1}{M}\sum_{i=1}^{M} P_i P_i^T \tag{11}$$

and

$$C_{VP} = \frac{1}{M}\sum_{i=1}^{M} P_i^T P_i \,. \tag{12}$$

Then we calculate

$U = [U_1, U_2, ..., U_l]$

and

$V = [V_1, V_2, ..., V_r]$,

where, $U_1, U_2, ..., U_l$ are the l eigenvectors respectively corresponding to the l largest eigenvalues of C_{UP}, and $V_1, V_2, ..., V_r$ are the r eigenvectors respectively corresponding to the r largest eigenvalues of C_{VP}.

Finally, bilateral projection is formulated as follows:

$$Y_i = U^T P_i V \,. \tag{13}$$

Getting the left multiplying projection matrix U and right multiplying projection matrix V respectively, we can obtain the feature matrix Y of any probe color sample image A by the bilateral projection mentioned above. After that, the Nearest Neighbor Classification can be employed to perform classification through the comparison between Y and $Y_i \,(1 \le i \le M)$.

IV. EXPERIMENTAL RESULTS

A. Face image database

CVL face database is made up of 114 people face images, and the 110 images of them were used in our experiments. We use the three front view images of each person as the experimental images, and randomly choose one of the three as probe sample and the other two remaining images as the training samples. Figure 2 presents the normalized images. The recognition rates were the average values of 500 runs of the same program.

Figure 2. Partial sample images in CVL human face database

FEI face database contains 200 people, and a subset of it was used in our experiments. The subset is made up of 200 people, with 2 front view pictures of each person (The expression of one picture is neutral, while another is smiling). Figure 3 shows some normalized images. In our experiments, we choose one picture of every person as probe sample randomly and the other picture as the training sample. Also, the recognition rates were the average values of 500 runs of the same program.

Figure 3. Partial sample images in FEI human face database

B. Experimental results and analysis

TABLE I. Experimental results of color image recognition based on method in [11]

Number of eigenvectors (d)	Recognition accuracy in FEI database(%)	Recognition accuracy in CVL database(%)
$d = 10$	76.5	66.36
$d = 20$	80.50	73.64
$d = 30$	82.50	78.18
$d = 40$	84.50	78.18
$d = 50$	84.50	79.09
$d = 60$	85.50	79.18
$d = 70$	87.00	80.00
$d = 80$	88.00	80.00
$d = 90$	88.50	81.82

TABLE II. Experimental results of gray image recognition based on NIB2DPCA in [6]

Number of eigenvectors		Recognition accuracy in FEI database(%)	Recognition accuracy in CVL database(%)
Line (l)	*Row(r)*		
$l = 1$	$r = 10$	22.00	41.82
$l = 1$	$r = 20$	29.50	47.27
$l = 1$	$r = 40$	33.50	48.18
$l = 1$	$r = 40$	34.50	52.73
$l = 2$	$r = 40$	52.50	67.27
$l = 3$	$r = 40$	69.50	70.91
$l = 4$	$r = 40$	71.50	76.36
$l = 5$	$r = 40$	79.00	74.55

TABLE III. Experimental results of our color image feature extracting method

Number of line eigenvectors (l_0)	Number of row eigenvectors (r_0)	Recognition accuracy in FEI database(%)	Recognition accuracy in CVL database(%)
$l_0 = 1$	$r_0 = 20$	89.05	75.45
$l_0 = 1$	$r_0 = 40$	92.00	80.91
$l_0 = 1$	$r_0 = 60$	93.00	82.73
$l_0 = 1$	$r_0 = 80$	93.50	84.55
$l_0 = 2$	$r_0 = 20$	90.50	82.73

$l_0 = 2$	$r_0 = 40$	93.00	88.18
$l_0 = 2$	$r_0 = 60$	93.50	89.09
$l_0 = 2$	$r_0 = 80$	93.50	90.00
$l_0 = 3$	$r_0 = 20$	92.00	83.64
$l_0 = 3$	$r_0 = 40$	95.00	89.09
$l_0 = 3$	$r_0 = 60$	95.00	91.82
$l_0 = 3$	$r_0 = 80$	95.50	91.82

Note. In table III, for FEI database the optimal numbers of both the line and row eigenvector are $l = 5$ and $r = 40$ respectively in feature pre-extraction step, while for CVL database the optimal numbers are $l = 4$, $r = 25$ respectively in the same step. l_0 and r_0 respectively denote the numbers of line eigenvectors and of row eigenvectors respectively in the feature extraction step.

TABLE IV. Comparison of our method with other two methods on both CVL and FEI human face databases

Face data base	Comparison items	Our color image feature extracting method	Gray image method in [6]	Color image method in [11]
FEI	*Recognition accuracy (%)*	95.50	79.00	88.50
	Training time (s)	1.287	0.0835	676.93
	Recognition time(s)	0.3688	0.1761	0.3421
	Total time (s)	1.6561	0.2596	677.2747
	Training sample number/class	1		
CVL	*Recognition accuracy (%)*	91.82	76.36	81.82
	Training time (s)	1.3695	0.1077	684.5592
	Recognition time (s)	0.4018	0.1983	0.3092
	Total time (s)	1.7713	0.3063	684.8684
	Training sample number/class	2		

Note. In table IV, the results were obtained under the optimal l , r , l_0 and r_0 .

Table Ⅰ and table III show what we have improved on method in [11]. Firstly, we used the NIB2DPCA to replace the 2DPCA so that the recognition accuracy increased; secondly, we added a feature pre-extraction step before final feature extraction that leads to the largely decrease of the size of the output matrix in this step. Table IV further shows that the recognition accuracy of our method increases by 7.00% to 10.00%.

In table II and table III, we can see that our method well keeps the color information contained in the correlations both within and between the R, G, B layers, while method in [6] keeps the color information contained in the correlations only between the layers. That leads to the increase of the recognition accuracy of our method. Table IV shows that, compared to the method in [6], the recognition accuracy of our method run on FEI and CVL color face databases increases by 16.50% and 15.46% respectively.

V. CONCLUSIONS

NIB2DPCA-based feature extraction method for color image recognition was proposed which can improve recognition accuracy. The reason why these advantages can be achieved is that the NIB2DPCA is effectively twice employed in our method. With the first use of NIB2DPCA in feature pre-extraction step, the image data are greatly compressed and the color correlations within and among the R, G, B layers are well kept. With the second use of NIB2DPCA in feature extraction step, the image data are further compressed and the color correlations are almost fully extracted into the feature matrix. Our method can be used in the computing environment with only limited computing ability and limited memory space, especially in embedded system.

REFERENCES

[1] Yip A, Sinha P. Role of color in face recognition[C].MIT tech report(ai.mit.com)AIM-2001-035 CBCL-212, 2001.

[2] Shih P, Liu C. Improving the face recognition grand challenge baseline performance using color configurations across color spaces[C]//IEEE Int'l Conf on Image Processing, 2006 : 1001-1004.

[3] Choi Y, Ro Y M, Plataniotis K N. Color face recognition for degraded face images[J]. IEEE Transactions on System Man and Cybernetics Parts B, 2009, 39(5) : 1217-1230.

[4] Turk M, Pentland A. Eigenfaces for recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1) : 71-86.

[5] Yang Jian, Zhang D, Frangi A F, et a1. Two-dimensional PCA : A new approach to appearance-based face representation and recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(1) : 131-137.

[6] GUAN Ye-peng.Robust video foreground segmentation and face recognition[J].Journal of Shanghai University (English Edition),, 2009, 13(4) : 311-315.

[7] PAN Quan , ZHANG Min-gui , Zhou De-long , et al. Face recognition based on singular-value feature vectors[J].Optical Engineering, 2003, 42(8) : 2368-2374.

[8] Torres L, Reutter J Y, Lorente L. The importance of the color information in face recognition[C]//Proc Int'l Conference on Systems, Man and Cybernatics, 1999, 3 : 627- 631.

[9] Rajapakse M, Tan J, Rajapakse J. Color channel encoding with NMF for face recognition[C]//International Conference on Image Processing, 2004 : 2007-2010.

[10] HUANG Xiao-hua, WANG Chun-mao, ZHENG Wen-ming. An information fusion and recognition method for color face images[J]. Journal of Image and Graphics , 2010 , 15(3) : 422-428. (in Chinese)

[11] WANG Cheng-zhang, YIN Bao-cai, BAI Xiao-ming, et al. Color face recognition based on 2DPCA[C]. Proceedings of the 19th International Conference on Pattern Recognition, Tampa, Florida USA, 2008.

3D Multi-Modality Medical Image Registration based on Quantum-behaved Particle Swarm Optimization Algorithm

Lihui
Jiangnan university
Network Information&Operation Center
Wuxi, China
e-mail:cwlh@jiangnan.edu.cn

Zhu Zhijun
Jiangnan university
Archives
Wuxi, China
e-mail: zzj@jiangnan.edu.cn

Abstract—**Image registration based on mutual information is of high accuracy and robustness.Unfortunately, the mutual information function is generally not a smooth function but one containing many local maxima, which has a large influence on optimization. This paper proposes a registration method based on Quantum-behaved Particle Swarm Optimization Algorithm. Not only QPSO has less parameters to con- troll, but also does its sampling space at each iteration covers the whole solution space.Thus QPSO can find the best solution quickly and guarantee to be global convergent. Experiments shows that this registration method could efficiently restrain local maxima of mutual information function and it can improve accuracy .Compare with the gold standard, the subvoxel accuracy can be achieved.**

Keywords-component; mage registration; Mutual information;Quantum-behaved Particle Swarm Optimization Algorithm

I. INTRODUCTION

Image registration is one of image processing technology developed rapidly in recent years. It is widely used in medical imaging technology, pattern recognition, artificial intelligence and electronic systems, mainly to the difference between the probe and the detection of image. In clinical diagnosis, radiation treatment planning and image guided surgery, often require the patient to accept a variety of computer tomography imaging, to provide the pathological and anatomical aspects of complementary information; but due to in imaging the patient positioning differences, and images of different resolution, different than the set of parameters to, the doctor is difficult only by imagination will be multi frame image accurately aligned, this time on the need for two or more medical images registration. Medical image registration is a kind of space transformation for a medical image, which makes it consistent with the corresponding points on the other medical images.This agreement is that the same anatomical points on the human body have the same spatial position on the two images.The fusion doctors can obtain the complementary information of the pathology and anatomy by image registration.

Image registration method based on mutual information although has high precision, the image pretreatment, and other advantages, but due to the interpolation calculation or there may be a better local matching leads to the goal function has a large number of local extreme values. This adds to the difficulty of many optimization algorithms. Through the introduction of QPSO algorithm, QPSO not only fewer number of parameters, the every Diego walking sampling space can cover the whole solution space, so as to ensure the global convergence of the algorithm. The other algorithm can easily fall into the local optimum, and the accuracy of the registration algorithm is improved, and the registration efficiency is reached 100%.

II. KEY TECHNOLOGY PROFILE

A. Mutual information

The mutual information is a basic concept of information theory, and is used to describe the statistical correlation between two random variables, and it is a measure of how much information is contained in a variable that contains the amount of information of another variable. It can be described by entropy:

$$I(A,B) = H(A) + H(B) - H(A,B) \quad (1)$$

Among them H(A) and H(B) Respectively image A and B entropy,H(A,B) the joint entropy of the two. In the multi mode image registration, when the spatial position of the two images is exactly the same, the information of the other images is expressed in one image, namely mutual information I (A, B) for the maximum.

Because the mutual information is sensitive to the change of overlapping area,Studholme[8] and Maes[9] two forms of normalized mutual information are proposed:

$$I(A,B) = \frac{H(A)+H(B)}{H(A,B)} \quad (2)$$

$$ECC(A,B) = \frac{2I(A,B)}{H(A)+H(B)} \quad (3)$$

Normalized mutual information can better reflect the change of registration function.

B. Three linear PV algorithm interpolation

In the registration process, due to the coordinates of the floating image by spatial transform obtained after is not necessarily an integer and need to by the interpolation method to get point transformation of gray value, floating the pixels of the map f from the sample a in a spatial transformation corresponding to the reference graph R B, usually B space coordinate with arbitrary a practical reference images do not overlap. Three linear PV

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.85

interpolation algorithm not through the neighbor points determine B gray, but according to the eight surrounding pixels and B space distance weight distribution, so that around 8 pixels contribution to the joint intensity distribution statistics, namely

$$\forall i : h(f, r(i)) += W_i, \text{且} \sum_i W_i = 1, \quad i = 1,2,\cdots,8 \tag{4}$$

Among them, R (I) is the gray level of 8 neighbors, Wi is the weight.

This interpolation method can make the calculation of mutual information more accurate, and it can be relieved for the optimization of local extreme problems.

III. QPSO ALGORITHM

QPSO[1][2] It is also a particle swarm evolution algorithm, which is also operated by the adaptive value of individual (particle). QPSO will each individual as N_d dimension a particle of weight and volume in search space, and in the search space to a certain speed of flight. The flight speed of the flight is adjusted dynamically by the individual and the group of flight experience. Each particle represents a position in the N_d dimensional space. Adjust the position of the particles in two directions:①The optimal position of each particle found so far;②Optimal position of particle swarm.

Each particle contains the following information:

(1) $x_i = (x_{i1}, x_{i2}, \cdots x_{id})$: Current position of particles;

(2) $v_i = (v_{i1}, v_{i2}, \cdots v_{id})$: Current velocity of particles;

(3) $P_i = (P_{i1}, P_{i2}, \cdots P_{id})$: Optimal value of I particle,That is pbest;

(4) $P_g = (P_{g1}, P_{g2}, \cdots P_{gd})$: Optimal adaptive value of particle swarm,That is gbest。

The evolutionary formula of the particles is:

$$mbest = 1/M \sum_{i=1}^{M} P_i = (1/M \sum_{i=1}^{M} P_{i1,} \cdots, 1/M \sum_{i=1}^{M} P_{id}) \tag{5}$$

$$P_{id} = \phi \times P_{id} + (1-\phi) \times P_{gd} \qquad \phi = rand \tag{6}$$

$$x_{id} = P_{id} \pm \alpha \times \left| mbest_d - x_{id} \right| \times \ln(1/u) \qquad u = rand \tag{7}$$

Among them,mbest is the intermediate position of the pbest particle swarm; P_{id} random points between P_{id} and P_{gd}, α is QPSO's coefficient of contraction,It's QPSO important parameter for convergence, general desirability $\alpha = (1.0 - 0.5) \times (MAXITER - T) / MAXITER + 0.5$ .Among them, T is the current number of iterations, $MAXITER$ is the maximum number of allowable iterations.

IV. ASSESSMENT OF REGISTRATION RESULTS

There is usually no so-called gold standard in medical image registration. However, the standard results can be obtained by a prospective and marker based registration method. Some patients in the Department of neurosurgery operation at the Vanderbilt University Medical Center, the data collection of the skull is fixed and received by the multi mode medical image (CT, MR, PET). The gold standard for the evaluation of the algorithm is obtained by the registration and localization markers. "

The researchers of the registration algorithm use the 3D multi-mode image data which has been erased mark points, after registration, the results will be submitted to the University of Vanderbilt for evaluation. Prior to the assessment, some interested areas (generally 10) were given by medical experts, often in the sensitive areas of Department of neurosurgery operation. Each ROI is defined in a MR image, while its center c is calculated; Then the gold standard of the prospective registration algorithm is determined and its corresponding point c' on PET are determined; Then the registration results of the algorithm are used to determine the corresponding point c'' of c' in MR; By calculating the distance between each origin c and the point c'', Registration error as target（Target Registration Error, Abbreviation TRE ）, The accuracy of the corresponding registration algorithm is also statistically. Figure 1 Schematic diagram for the process, c around the region of the region of interest to the region of MR; c is center, c' is the corresponding points on the PET image obtained by the application of the gold standard inverse transform, $c' = G^{-1}(c)$; $c'' = R(c')$, c'' is applied to the registration algorithm to evaluate the transformation results obtained from the corresponding points of the c' on the MR image. The geometric distance d of the corresponding point c'' after the origin c and the two transformation is calculated, and the error of the registration algorithm can be determined.

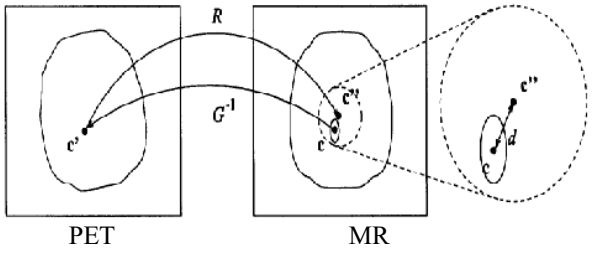


Figure 1 Methods of accuracy evaluation for the image registration algorithm

The coordinates of the coordinates of the 3D image before the registration are shown in Figure 2.Among them(a) is floating image;(b) is reference image. The coordinates of the 8 points are the center of the 8 vertices.

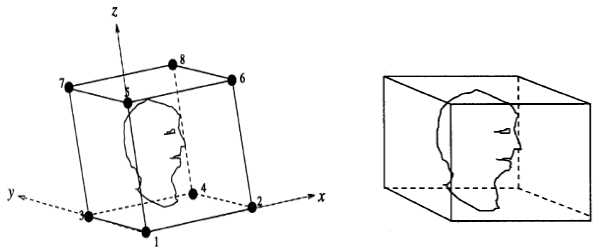


Figure 2 three dimensional image before registration:(a)is loating image;(b)is reference image

After registration the coordinates of the 3D image and 8 points are shown in Figure 3. The coordinates of the 8 points are the center of the 8 vertices. The points between $1'$ and 8 are points after the space coordinate transformation from $1'$ to 8.

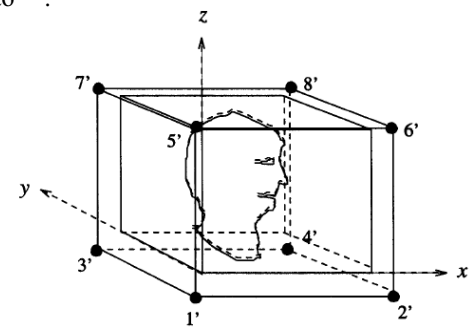


Figure 3 Three dimensional image after registration

The retrospective image registration algorithm evaluation project is a kind of "double blind" research process.. The so-called double blind, that the assessment staff do not know the specific algorithm is evaluated, and the algorithm researchers do not know the gold standard until all registration results submitted. This makes the assessment of the algorithm more true and reliable, and it also accords with clinical practice.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This paper uses the image data from the University of Tennessee, Tennessee Vanderbilt retrospective image registration algorithm assessment project, The project is supported by the United States NIH, number is 8R01EB002124-03.Professor J.MichaelFitzpatrick is the principal person in charge of the project. In this experiment, We use 35 sets of PET-MR images provided by Vanderbilt to experiment. We evaluate the results of 35 sets of PET-MR image registration sent to Vanderbilt University. Using the proposed registration method, PET-MRI image registration, the MR image as the reference image, PET images as floating images. We evaluate the results of 35 sets of PET-MR image registration sent to Vanderbilt University. The results of the assessment are shown in Table 1. The results of the assessment can be seen in the following website http://insight-journal.org/rire/view_results.php，The name is Jun Sun。

TABLE I. THE ERROR OF OUR REGISTRATION RESULTS WITH THE GOLD STANDARD ERROR (PET-MRI) (UNIT MM)

	PET-PD	PET-PD_r	PET-T1	PET-T1_r	PET-T2	PET-T2_r
mean	3.107	2.715	2.798	1.913	2.831	2.136
median	2.929	2.61	2.387	1.528	2.515	1.548
maximum	5.292	4.585	6.507	4.551	5.78	4.65

Taking the pixel diagonal distance of PET image as a pixel size, That is：

$$\sqrt{2.590723^2 2.590723^2 + 8.000000^2} \approx 8.799(mm)$$

Below the registration results with the gold standard website(//insightjournal.org/rire/view_results.php http:) on Hahn Dieter (Erlangen University Nuremberg January 2009 registration results, as well as Rohlfing Torsten(International SRI) registration results of December 2009 were compared. Comparison.Results as table 2~ 4.

TABLE II. MEAN ERROR

	Dieter	Torsten	This paper
PET-PD	3.824	3.627	3.107
PET-PD_rect	3.439	2.822	2.715
PET-T1	3.240	3.174	2.798
PET-T1-rect	3.173	2.330	1.913
PET-T2	2.717	3.156	2.831
PET-T2_rect	3.495	3.051	2.136

TABLE III. MEDIAN ERROR

	Dieter	Torsten	This paper
PET-PD	3.855	3.507	2.929
PET-PD_rect	3.294	2.274	2.61
PET-T1	3.117	3.154	2.387
PET-T1-rect	2.920	1.966	1.528
PET-T2	2.575	3.231	2.515
PET-T2_rect	2.641	2.473	1.548

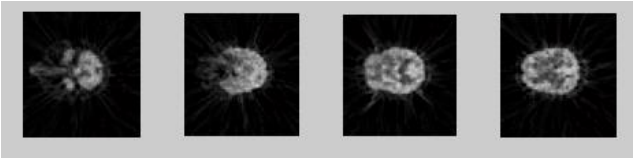
TABLE IV. MAXIMUM ERROR

	Dieter	Torsten	This paper
PET-PD	3.824	3.627	3.107
PET-PD_rect	3.439	2.822	2.715
PET-T1	3.240	3.174	2.798
PET-T1-rect	3.173	2.330	1.913
PET-T2	2.717	3.156	2.831
PET-T2_rect	3.495	3.051	2.136

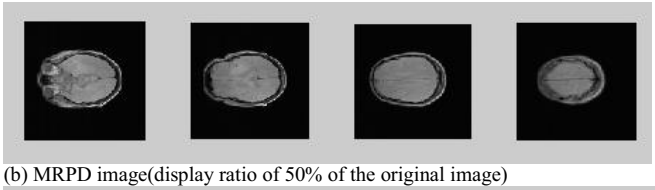
From table 1 can see, Medical image registration method based on QPSO algorithm, A total of 35 sets of registration data of the PET-MR 3D body registration data provided by Vanderbilt university are evaluated for our results, Compared with the "gold standard", the error mean, the median and the maximum value are the registration accuracy of sub-pixel. At the same time from table 2, table 3 and table 4 can see that our method with the gold standard online other methods to

compare the accuracy of higher than most other methods. Although the "gold standard" website other registration algorithm with individual error is less than the error, but these algorithms cannot guarantee mean, median and maximum value of the error at the same time to achieve sub-pixel registration accuracy. And our algorithm can guarantee the error of mean, median and maximum value of the sub pixel registration accuracy.

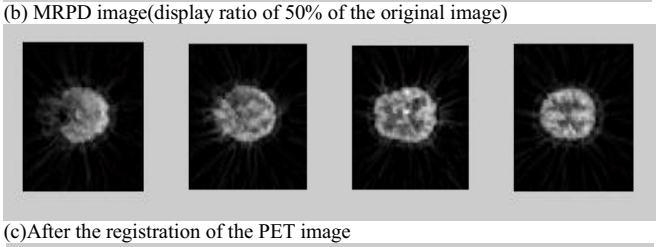
Fig. 4 results of the registration of the PET-MRPD images of the patient 002. Figure (a) is PET image, figure (b) is MRPD images, from left to right were first, third, sixth and ninth layers. The contrast graph (a), graph (b) can know the space position difference of two images is far. Figure (c) after registration of the PET image. In order to facilitate visual, we extract mrpd image edge, and added to the PET images, as shown in the diagram (d), at all levels of the edges, two images reached good registration.



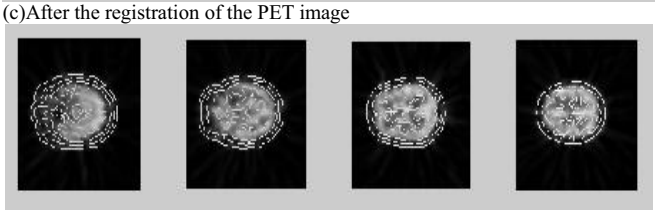
(a) Original PET image



(b) MRPD image(display ratio of 50% of the original image)



(c)After the registration of the PET image



(d)The MRPD edge added (c) extracted from (b) is applied to the results of the PET images shown

Figure 4 The registration results of PET and MRPD (from left to right are 4, third, sixth and ninth).

VI. CONCLUDING REMARKS

Based on the mutual information registration method has high precision, strong robustness, does not require pretreatment, and other advantages, but there are also optimization for a long time, mutual information function has many local minima, the commonly used some optimization algorithm, such as the Powell method easy to fall into local optimal solution, can not be globally optimal registration result. This paper presents a on QPSO optimization algorithm as a search strategy, with mutual information as similarity measure, the 3D PET, MR image registration, good results have been obtained, the accuracy of image registration results reached sub-pixel level, as compared with other method is better than other method is better. In the fast development of imaging technology, the image types are more and more today, and the method of high performance registration is very important for multi mode image registration.

REFERENCES

[1] SUN Jun, XU Wenbo. A global search strategy of quantum-behaved particle swarm optimization: proceedings of IEEE Conference on Cybernetics and Intelligent Systems[C]. [S. l.] : [s. n.] , 2004: 1112116.

[2] SUN Jun, FENG Bin, XU Wenbo. Particle swarm optimization with particles having quantum behavior: p roceedings of 2004 Congress on Evolutionary Computation[C]. [S. l.] : [s. n.] , 2004: 3252331.

[3] Klein S, Staring M, Pluim J.P.W. Evaluation of Optimization Methods for Nonrigid Medical Image Registration Using Mutual Information and B-Splines[J]. Image Processing,IEEE Transactions Volume 16(12) Dec. 2007:2879-2890.

[4] Yutaro Yamamura; Hyoungseop Kim; Akiyoshi Yamamoto. A Method for Image Registration by Maximization of Mutual Information. SICE-ICASE,2006.International Joint Conference Oct. 2006:1469-1472.

[5] Seok Lee, Minseok Choi, Hyungmin Kim, Park, F.C. Geometric Direct Search Algorithms for Image Registration[J]. Image Processing,IEEE Transactions Volume 16(9) Sept. 2007:2215 – 2224.

[6] Ximiao Cao, Qiuqi Ruan. A Survey on Evaluation Methods for Medical Image Registration. Complex Medical Engineering,2007.CME2007.IEEE/ICME International Conference May 2007:718 – 721.

[7] Wachowiak M P, Smolikova R, and Zheng Y. An approach to multimodal biomedical image registration utilizing particle swarm optimization. *IEEE Trans. on Evolutionary Computation*, 2004, 8(3): 289–301.

[8] Studholme C, Hill D L G, and Hawkes D J. An overlap invariant entropy measure of 3D medical image alignment.*Pattern Recognition*, 1999, 32(1): 71–86.

[9] Maes F, Collignon A, and Vandermeulen D, *et al*..Multimodality image registration by maxim- ization of mutual information. *IEEE Trans. on Medical Imaging*, 1997, 16(2):187–198.

[10] Luo Shu-qian,Zhou Guo-hong,Medical image processing and analysis[M]. Science Press ,2003:140-202.

[11] Yang Fan,Zhang Han-ling.Multiresolution 3D Image Registration Using Hybrid Ant Colony Algorithm and Powell's Method [J]. Journal of Electronics & Information Technology,2007,29(3):622-625.

[12] Wang Anna,Wu Jie,Zhang Xinhua,Tao Ran. A Novel Medical Image Registration Algorithm Based on PSO and Wavelet Transformation Combined with 2v-SVM.Innovative Computing, Information and Control,2007.ICICIC07.Second International Conference Sept. 2007:584-

Research on medical image registration based on QPSO and Powell algorithm

PAN Ting-ting
Wuxi City College of Vocational Technology
College of things Engineering
Wuxi, China
e-mail:panti1984@163.com

ZHAO Ji
Wuxi City College of Vocational Technology
College of things Engineering
Wuxi, China
e-mail: 4293713@qq.com

Abstract—**This paper proposed a multi-resolution search optimization algorithmcombining QPSO and Powell algorithm,it can solve the Powell algorithm shortcomingeffectively. In this paper, the algorithm were usedin two-dimensional MRI images registration.The experimental results show that the algorithm can effectively overcome the problem of local extremum of mutual information function, and improve the precision and speed of registration.**

Key words-***Image registration; Mutual information; QPSOalgorithm; Powell's method***

I. INTRODUCTION

To obtain comprehensive information in many aspects of the patients though analysising several images of the same patient together in medical image analysis,that improve the level of medical diagnosis and treatment. For the quantitative analysis of different images ,we must first solve the strictalignment problem of that few images, called the image registration.

Medical image registration is to seek a (or a series of) space transformation for medical images, that make it reach space consistent with the corresponding points on the other medical images. This consistency refers to the same anatomical point of the body has the same spatial position in two pieces of matching images.

II. IMAGE REGISTRATION ALGORITHM

A. Powell algorithm

Powell algorithm is a kind of pattern search method. It is the best direct search method, and has strong local search ability, that only need to compute the objective function value without its derivative value, but existing local extremum problems. In order to solve the disadvantages of the Powell algorithm, this paper proposed QPSO algorithm for global optimization, and then combine the Powell algorithm for image registration.

B. mutual information registration principle

Mutual information is used to describe the statistical correlation between two random variables, is refer to a measure of how muchinformation a variable contains another variable. In multimode image registration, when the spatial position of the two images is completely consistent, the expression in the image of another image information,which is the largest of the mutual information .

Because of the mutual information is more sensitive to the change of overlap, Studholme and Maes put forward two kinds of normalized mutual information form, that can reflect the change of the registration function better.

C. QPSO algorithm

QPSO is also a kind of particle swarm evolution algorithm, using the concept of group and evolution, but also operating on the adaptive value of the individual (particles) size .

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.86

Each individual is seemed a particle without weight and volume that flighting at a certain speed in N_d dimension of search space in QPSO. The flight speed is adjusted dynamically by flight experience of individual and group.Each particle represents a location in N_d dimensional space, towards the adjustment of the particle following two directions: (1) the optimum position of each particle found so far; (2) the optimal position of the particle swarm.

Each particle contains the following information:(1) The current position of the particle;(2)The current velocity of particle;(3)The best adaptive of particle i, namely pbest;(4) The optimal value of the adaptive particles warm, namely gbest。

III. THE ALGORITHM OF THIS PAPER

A. remove the background

In order to make the image interference noise free, we first remove the background of the image. Remove the image background using the method proposed by Wan Rui. The specific steps are as follows:

(1)Calculate the maximum and minimum of gray level in the image and set the initial value of the threshold:

$$T_0 = \frac{Z_1 + Z_k}{2}$$

(2)Divide the image into two parts according to the threshold, calculate the average gray value of the two part(Z_0 and Z_B):

$$Z_0 = \frac{\sum_{z(i,j)<T_k} z(i,j) \times N(i,j)}{\sum_{z(i,j)<T_k} N(i,j)}$$

$$Z_B = \frac{\sum_{z(i,j)>T_k} z(i,j) \times N(i,j)}{\sum_{z(i,j)>T_k} N(i,j)}$$

In the formula, Z(i,j) is the gray value of point (i,j) in the image, N(i,j) is the weight coefficient of point(i,j), here N(i,j)=1.0.

(3)Calculate the new threshold

$$T_{k+1} = \frac{Z_0 + Z_B}{2}$$

If $T_k = T_{k+1}$,the end. Otherwise $k = k + 1$, the iteration to perform the above steps.

Finally remove the background using seed filling method from the upper left corner of the image to start filling point that is less than the threshold.

B. Interpolation

The point in floating images is not necessarily an integer through space transformation, it need the gray value of transformation point by interpolation method. Trilinear Partial Volume distribution in interpolation is often used in 3D image registration based on mutual information. PV interpolation algorithm does not introduce new gray value, contribution of the gray f (P) of floating point P in the image of the joint histogram is 8 points nearest around point Q in the image and the same three linear interpolation algorithm and weighted. 2D images use the similar method.

C. Point out strategy

When a sample points P_f in the floating image f after the T space transformation, corresponding point falls outside of the reference map r, called points out point. Obviously, the calculation of mutual information must be considered out of bounds point. Some scholars could ignore out points, namely using different sampling points to calculate the mutual information in different iteration cycle, or set these points out gray approximate to zero, the experimental results show that has bad effects on registration accuracy.

In the experiment we use the out point strategy proposed by LuoShuqian, it

is to point out the gray that is equal to its nearest boundary pixel. This expand the background of the reference map , and keep the number of samples the same in the process of optimizing , so the calculation of mutual information is more accurate.

D. registration optimization algorithm on QPSO algorithm and Powell method

We combine global optimization ability of QPSO and local optimization ability of Powell well in the process of registration optimization. Since the data in the original image is very large,thus QPSO algorithm to optimize time longer. We use the multi-resolution method based on wavelet transform, optimize by the PV interpolation, and image gray series is set to 256. The optimization process is divided into two steps: step 1, using QPSO algorithm for registration in a low resolution image,the image data quantity is small, the mutual information calculation speed and the optimization process can be completed quickly; Step 2, set the initial point of Powell method as the best solution of QPSO, using Powell method in the high resolution image optimization.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This paper selects one MRI image (256 x 256) to experiment, specific steps: first,remove background of the image,and then rotate the image after removing background clockwise for 30 °,and pan the image down 5 pixels and right 5 pixels.It seemed the original image that removing background as a reference image,and the image after the transformation as a floating image.There is four kinds of algorithms for registration, each algorithm run ten times, respectively, the registration results and time as shown in table 1.

TABLE 1 REGISTRATION RESULTS AND RUNNING TIME OF MRI IMAGE 1

algorithm	$\Delta\theta$ RMS	Δt_x RMS	Δt_y RMS	TMRT
Powell	0.5564	0.8837	0.9486	352
PSO	0.1025	0.5133	0.4755	17268
QPSO	0.0050	0.1115	0.1699	16925
QPSO+Powell	0.0048	0.1167	0.1815	1253

Reference image is shown in figure 1, reference image after removing background is shown in figure 2, the floating image is shown in figure 3. In the table, $\Delta\theta$ is the rotation Angle error (unit is angle), Δt_x is the error of the X axis translation,and Δt_y is the error of Y axis translation (unit is pixel) .T is the time used in the algorithm,RMS is root - mean – square of each set of data, and MRT is the average running time. The number of iterations of QPSO and PSO is 300 times in experiments. The QPSO+Powell method has two levels of wavelet decomposition for the original image.The image after two levels of wavelet decomposition is shown in figure 4,floating image is shown in figure 5. Experiments were completed under the Matlab6.5, running environment for the processor Pentium4, 2.0MHz, memory 256MB.

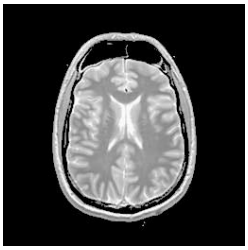


Fig.1 Reference image

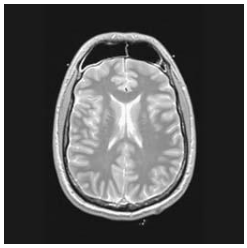


Fig.2 reference image

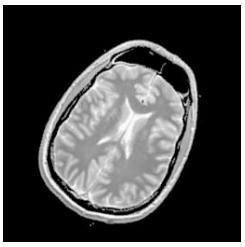


Fig.3 the floating image

Fig.4 reference image after two levels of wavelet

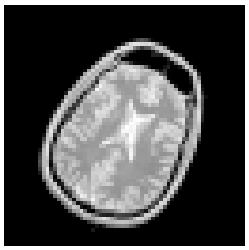


Fig.5 floating image after two levels of wavelet decomposition

The initial parameter of Powell method is (0,0,0), but the initial parameter of PSO and QPSO is randomly generated. In this case, the error rate of various algorithms is shown in Table 2.

TABLE2 THE ERROR RATE OF VARIOUS ALGORITHMS

algorithms	the error rate（%）
Powell	50
PSO	20
QPSO	0
QPSO+Powell	0

V.CONCLUSION

The registration method based on mutual informationhas much advantages: high precision ,strong robustness,not need to extract image features and so on.But there're some problems : optimized time is long, mutual information function has many local minima.It's easy to fall into local optimal solution using the commonly methods(such as PSO and Powell),that can't get the global optimal solution. .A multi resolution search optimization algorithm based on QPSO and Powell is proposed in this paper. This method use multi-resolution strategy based on wavelet transform, using mutual information as similarity measure, the registration of 2D MR images by combining the QPSO algorithm and Powell method and good results have been obtained. The accuracy of image registration results is very high. The method can also be used for registration of multi-mode images and three-dimensional image registration.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from "Qing LanProject[2012-16]" of Jiangsu province and National Natural Science Foundation of China 61300149.

REFERENCE

[1] SUN Jun, XU Wenbo. A global search strategy of quantum-behavedparticle swarm optimization: proceedings of IEEE Conference on Cybernetics and Intelligent Systems[C]. [S. l.] : [s. n.] , 2004: 1112116.

[2] SUN Jun, FENG Bin, XU Wenbo. Particle swarm optimization withparticles having quantum behavior: p roceedings of 2004 Congress onEvolutionary Computation[C]. [S. l.] : [s. n.] , 2004: 3252331.

[3] Maes F, Vandermeulen D, and Suetens P. Comparativeevaluation of multiresolution optimization strategies formultimodality image registration by maximization of mutualinformation. Medical Image Analysis, 1999, 3(4): 373–386.

[4] Pluim J P W, Maintz J B A, and Viergever M A. Mutualinformation based registration of medical images: A survey. IEEE Trans. on Medical Imaging, 2003, 22(8): 986–1004.

[5] Plattard D, Soret M, and Troccaz J, et al.. Patient set-upusing portal images: 2D/2D image registration using mutualinformation. Computer Aided Surgery, 2000, 5(4): 246–262.

[6] Jenkinson M and Smith S. A global optimization method forrobust affine registration of brain images. Medical ImageAnalysis, 2001, 5(2): 143–156.

[7] Wachowiak M P, Smolikova R, and Zheng Y. An approach tomultimodal biomedical image registration utilizing particle swarm optimization. IEEE Trans. on EvolutionaryComputation, 2004, 8(3): 289–301.

Face Tracking Algorithm based on Online random forests Detection

Fang Bao[1,2]
Jiangyin Polytechnic College [1]
Information Intelligence Fusion Research Engineering center of Jiangsu[2]
jiangyin Jiangsu china
515945108@qq.com

Yankai Zhang
Jiangnan University
wuxi Jiangsu china
472231859 @qq.com

***Abstract*—The Paper proposed a face tracking algorithm based on online random forests. The algorithm achieved detection-based tracking using online incremental extremely random forests detector, P-N learning is added to correct the detection error, and the dynamic target framework is proposed to maintain the online training set. The proposed algorithm integrated the results of the detector and the P-N learning, and the similarity to the dynamic target framework, thus the tracking position is confirmed. The experimental results show, the proposed algorithm could perform the tracking to any face rapidly and stably in a long-term period and complex background. Thus have practical worth in many application areas.**

Keywords- Online Learning; Extremely Random Forests; P-N Learning; Dynamic Target Framework; Face Tracking

I. INTRODUCTION

Under the condition of the widespread audio monitor, the rapidly recognize and tracking technical in the remote and uncooperative situation is acquired so much. Human face could perform the identity, alarm and tracking in the most natural way. Thus the subject of face detection, recognizing and tracking is a hot point in the machine vision area.

Moving face detection and tracking could be divided into static and dynamic background. In the situation of static background, AdaBoost[1,2] algorithm is often been used in detection, Mean-Shift[3] and Particle Swarm Filter[4] algorithm are often used in tracking. Face tracking under the static background could only been such short-period and limited tracking, often been in ticket check in.

In the internet dynamic background, selecting arbitrarily face, performing long-term tracking, is the most demanding application. Traditional off-line machine learning could not solve such problem, the on-line incremental learning pattern is demanded.

Now, decision tree is often used in incremental learning, especially the random forest[5,6]. But in both avid and statistical algorithm, the required data quantity is so huge, and the optimal attribute or statistical value should be recounted in the process of incremental restructure.

The Extremely Random Forest(ERF) proposed by Geutrs[7] et.al use the original sample as training set, the split threshold is selected randomly, the classify accuracy and time cost is prior to the traditional random forest. Thus the ERF is more suitable for small scale sample set. Face tracking in video sequence could been defined as a typical two classification problem with small scale sample set, so could be performed by the online incremental ERF algorithm.

The P-N learning[8] is often been used to correct the classify error. P-experts assign 'positive' to those negative sample defined by the classifier, but ought to be positive according to the constraint, and put it into the training set. N-experts assign 'negative' to those positive sample defined by the classifier, but ought to be negative according to the constraint, and put it into the training set.

This paper proposed a novel algorithm, we use the ERF classifier to perform the online incremental detection, the P-N learning is added to increase the detect precision, and the dynamic target framework is constructed online. The algorithm integrated the results of the detector, the P-N learning, and the similarity to the dynamic target framework, to confirm the tracking position online.

The following chapters are listed below. Section 2 introduces the modeling pattern of face and the dynamic target framework. Section 3 illustrates the online incremental detect based on ERF classifier. Section 4 is the integration of the classifier and P-N learning. Section 5 is the experiment result and corresponding analysis. The summary and further thinking is given in the last.

II. FACE MODELING AND THE DYNAMIC TARGET FRAMEWORK

A. *face modeling*

This paper use 2bitBP(2bit Binary Pattern) for face modeling. This is a Harr-Like[9,10] feature, there have only 4 kinds of codes through quantization. 2bitBP feature has good robustness to the light changing of the environments, but has high check fail rate in the self pose changing of face. The skin color pattern[11] is robust to the pose changing. So, we will use the YC_gC_r color space later to improving the robust to pose changing.

B. *the dynamic target framework*

We introduce a kind of dynamic target framework, it is a set of positive and negative samples, and is the training set. Let p^{i+} be the ith positive sample been added, and p^{i-} be the ith negative sample. Given arbitrary patch p and dynamic target framework M, we define concepts as below.

The similarity of patch, is the similarity between 2 patchs using Normalized Cross-Correlation(NCC).

$$S(p_i, p_j) = o.5 * (NCC(p_i, p_j) + 1) \quad (1)$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.87

The similarity of positive sample

$$S^{+}(p,M)=\max S(p,p_i^{+}) \quad (2)$$

The similarity of nearest neighbourhood positive sample

$$S^{r}=S^{+}/(S^{+}+S^{-}) \quad (3)$$

In video frame 1, the positive sample is the face to track. Then, within the range of 1 percent, making the geometric transformation 20 times by rotating and shifting in the face area, thus 20 positive samples are generated. Negative samples are generated outside the face area randomly. These labeled samples construct the original dynamic target framework, could be used to train the original ERF. In the following tracking process, positive and negative samples will be labeled and added into the framework, and an online training set is been maintained.

III. ONLINE INCREMENTAL FACE DETECTION BASED ON ERF

ERF consists of many decision trees, it use the original samples as training set, the split threshold is selected randomly at each tree's decision node. ERF proposed by Geutrs don't support online learning, thus WangAiPing et.al propose an ERF algorithm which support the online incremental learning, calls Incremental Extremely Random Forest(IREF) [12].

IREF store the new sample on the leaf node, whether the split is necessary on the node is decided by the Gini coefficient. Each decision tree is called incremental super tree. Let input sample be p, tree index be i and output incremental super tree be T_i, the construct process describes as below.

Stage 1:

If the root node of T_i not exists

return a incremental super tree t_i with root node R, R's label is the same as p's

Else

training sample p classifies to leaf node L

p is stored into L's sample list

update the count of sample's classification in L

if Gini(L)> split threshold

construct a new incremental super tree T(according to the steps in stage 2)

if L is the root node

let T to be the new root

else

let T to be the child node of the father node instead of L

endif

delete the leaf node L

return incremental super tree T_i

endif

Endif

Gini coefficient is used to judge the purity of the sample set. If the sample set D have k classifications, p_i is the proportion of category i, the calculate formulation of Gini coefficient is:

$$Gini(D)=1-\sum_{i=1}^{k}p_i^{2} \quad (4)$$

If the Gini coefficient in one node exceeds the threshold, the confusion degree of it is considered to be high enough, and ought to be split, then the incremental super tree is constructed using the samples stored on it.

Split threshold is formulated as below. Suppose there are k kinds of samples in certain leaf node, and the index of the largest number is m, the rate of other samples to mth is α_i, the sum is $\alpha=\sum_{i=1,i\neq m}^{k}\alpha_i$, then the threshold is calculated as:

$$\Delta=1-\left(1/(1+\alpha)\right)^2 \quad (5)$$

When we use the samples stored to construct the incremental super tree, let the input sample be S, the output tree be T, $n_{\min}$ be the minimum number of sample set. The step are list below as stage 2.

Stage 2:

If $|S|<n_{\min}$

or all the candidate characters in S don't change

or all the output in S are the same

generate a new leaf node T

store all the samples in S on T

count the distribution of all the classifications in S

return the leaf node T, it's label is decided by the distribution of all the classifications

Else

split S into 2 subsets named S_r,S_l, decision attribute and split test using s* of Geutrs in [11]

construct sub incremental super tree T_r,T_l respectively for Sr,Sl

according to s*, decision node T is generated, it's right and left sub tree is T_r,T_l

return the sub incremental super tree T

Endif

The initial training of ERF in our paper is performed when the first frame entered and the initial labeled target framework been formed. As the new frame enters, it is been handled using scan window, the size of the window is the same as the initial face patch. After normalization, each new patch is to be a new online incremental sample, and is labeled by the above algorithm. The final decision result is achieved by the average probability of all incremental super trees, so,whether it is the face sample is been decided.

IV. P-N LEARNING

When the face sample is detected by the above algorithm, we should put it as a given sample into the dynamic target framework. But there may exists check error in the detection, so we add the P-N Learning module to correct the check error.

By using the structural characteristic between the labeled and non-labeled samples, P-N Learning train and improve the classify performance of the classifier. The process is listed below.

Prepare a training set with few samples and a test set with much more samples. Train a formal classifier with the training set, and set up the transcendental constraint conditions. Label the test sample using trained classifier, found out those who are contradictory to the constraint

conditions. Re-label the samples who are contradictory using the constraint conditions, add them into the training set, thus the classifier is to be re-trained.

In our paper, appropriate constraint conditions should be ascertained, new sample been labeled by ERF classifier and are contradictory to the conditions should be re-labeled and put into the dynamic target framework, thus to improve the detector effects.

According to the special application area in our paper, the first constraint condition to be set is, the skin color pattern in YC_gC_r color space, this space is considered to be the best in the skin color clustering. The formulation from RGB color space to YC_gC_r color space is:

$$\begin{bmatrix} Y \\ C_g \\ C_r \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.4 & 128.6 & 24.9 \\ -81.1 & 112 & -30.9 \\ 112 & -93.8 & -18.2 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (6)$$

The scope belongs to face color, also the first constraint condition in our paper is:

$$C_g \in [85 \quad 135], C_r \in [-c_g + 260 \quad -C_g + 280] \quad (7)$$

When the overwhelming majority of pixels in one sample are belong to the above scope, and it is been labeled as negative by ERF classifier, then the sample should be re-labeled as positive and put into the dynamic target framework. Also, when the overwhelming majority of pixels in one sample are not belong to the above scope, and it is been labeled as positive by ERF classifier, then the sample should be re-labeled as negative.

Secondly, face tracking has particular characteristic in space domain, the target could only just be in one place in one video frame. So, when one sample is considered to be positive by both the ERF classifier and constraint condition, other samples only could be negative. This is the second constraint condition, called unique constraint.

In conclusion, our tracking algorithm is summarized as below.

In frame 1, arbitrary face is selected, the positive, negative samples and the initial dynamic target framework are generated, the initial ERF classifier is trained too. In the following frames, new samples is obtained by window scans, it is to be labeled by online incremental ERF, the P-N Learning is added to re-label it. At the end of window scan in each frame, if there exists samples which are fit to constraint condition 2, then the sample's position is the current tracking position. Else, all the similarity of positive sample of positive samples is calculated as formulation 4, the biggest one's position is to be the current tracking position.

V. EXPERIMENTAL RESULT AND ANALYSIS

In experimental segment, we draw in some important skills, and excellent tracking accuracy and speed is obtained.

Search only in the skin color connected area. When the frame is transferred to the YC_gC_r color space, found the space which fit the formulation $C_r \in [C_{r0} - 5, C_{r0} + 5]$, $C_g \in [C_{g0} - 5, C_{g0} + 5]$, generate all the connected area by the edge of each space, these space are to be the search space. So the search area is reduced greatly, which improve the tracking speed effectively. These connected search area are shown as white black area in each picture's up-left corner.

According to the change of face pose and size, we draw in the change frame coefficient[1,1;0.8,1;1.3,1;1,0.8;1,1.3], which correspond to the situation of not change, length narrow, length broaden, width narrow ,width broaden. Multiply the previous face patch with one coefficient, use the new patch to scan the skin color connected area. Thus, there are 5 screen shots for each patch. By the decision of ERF, the best screen shot is called the current best screen shot.

Judge the current best screen shot using skin color pattern constraint, if not, it is been put into the dynamic target framework as negative sample(shown as dotted box in below). And the following steps are skipped and the next video frame to be checked.

Select those whose NCC similarity of positive sample is positive from 5 screen shots, calculate their distance to the current best screen shot. If the distance is bigger than half of face size, then it is also been put into the dynamic target framework as negative sample. This step judge the patches using the unique constraint. For those screen shots which still have not been negated, use the ERF's decision as their weight, calculate the weighted mean of them, thus the weighted screen shot is obtained.

Finally, if the current best screen shot is been agreed by less than 95 percent of trees, then the weighted screen shot is been put into the dynamic target framework as positive sample. If the current best screen shot is been agreed by more than 95 percent of trees, then it is to be the current unique tracking target (all shown as solid line box in below).

The experiment result is shown below.

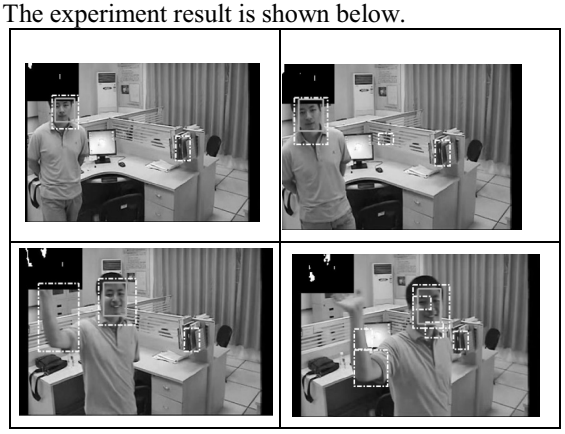


Figure 1. Picture set of complete algorithm

Pictures in figure1 show the face tracking effect using complete proposed algorithm. In the whole period from enter to leave of the face, the tracking is stable and quickly.

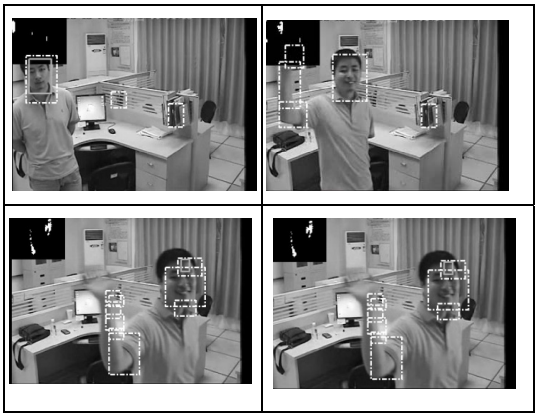


Figure 2. Picture set of algorithm without P-N Learning

Pictures in figuret2 show the tracking effect using algorithm without P-N Learning. We can see, without P-N Learning, the tracking could be so unstable.

Figure 3. Picture set of algorithm without NCC compare

Pictures in figure3 show the tracking effect using algorithm without NCC comparing. Without NCC comparing, the tracking speed is better, but the tracking stability is not so good as compared with the complete algorithm.

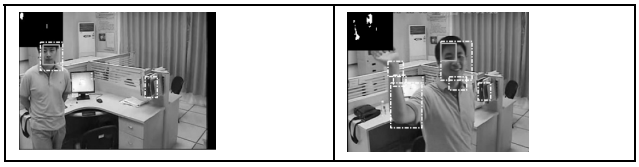


Figure 4. Picture set of algorithm without change frame

Pictures in figure4 show the tracking effect using algorithm without change frame coefficient. It leads to the locality of the target face.

Table1 show the average process time of each frame(second), numbers of frames to be processed per second and the correct tracking rate of al1 the referred algorithms.

TABLE I. COMPARE OF PROCESS SPEED AND ACCURACY

algorithm	average process time	frames processed	correct rate
Complete algorithm	0.0602	16.6	98.2%
without P-N	0.0452	22.1	76.5%
without NCC	0.0505	19.8	92.1%
without change frame coefficient	0.0413	24.2	89.4%

Integrally thinking about the speed and accuracy, the complete algorithm is the best. In the real world application, the real time process speed is required to achieve 25 frames per second. If get rid of the NCC or change frame coefficient, the tracking effect could be acceptable generally, and the tracking speed could reach the real time require easily.

VI. CONCLUSIONS

The Paper achieved detection-based tracking using online incremental extremely random forests detector, the P-N learning is added to improve the detection performance, and the dynamic target framework is proposed to construct the online training set. The proposed algorithm integrated the results of the detector and the P-N learning, and the similarity to the dynamic target framework, thus the tracking position is confirmed. Some economical skills like the skin color connected area, and the change frame coefficient are draw in, and the tracking effect is be enhanced obviously.

Experimental results show, the proposed algorithm could perform the tracking to any face rapidly and stably in a long-term period and complex background, thus could perform the practical level tracking.

REFERENCES

[1] Paul Viola,Micheal Jones.Rapid Object Detection using a Boosted Cascade of Simple Features[C]. Kauai Marriott,Hawaii, Proc of CVPR2001:511-518

[2] Yangfeng Deng, GuangDa Su, Bo Bo. A Rapid Dynamic Face Detection Algorithm Based on Adaboost [J]. Computer Enginnering , 2006, 32（11） : 222-224

[3] Juan Wang, JiaoMin Liu, JunYing Meng. Research of Modified Mean Shift Algorithm in MovingObject Tracking[J]。Journal of System Simulation, 2012, 24（9） : 1896-1899

[4] HaiTao Yao, XiFu Yao, HaiQiang Chen. A Self- Adapting tracking algorithm Based on PSO ParticleFiltering [J]。Wuhan University Journal of Natural Sciences, 2012, 37（4） : 492-495

[5] Leo Breiman.Random Forests[J]. Machine Learning:2001,45(1):5-32

[6] ZuHua Liu, HuiLin Xiong. Object Detection and Position Based on Random Forests[J]. Computer Enginnering, 2006,63(2):3-42

[7] .Pierre Geurts,Damien Ernst,Louis Wehenkel. Extremely Randomized Trees[J]. Machine Learning, 2006,63(2):3-42

[8] Zdenek Kalal,Tiri Matas,Krystian Mikolajczyk. P-N Learning:Boosrapping Binary Classifiers y Structural Constraints. VCPR2010 proceedings, 49-56

[9] YonmgFei Chen, XinMin Liu. Human Eye Detection Based on Skin Color and Harr-Like Feature. [J]. Computer engineering and Application, 2008,44（33） : 174-176

[10] YouDong Ding, XiaoFeng Du, XiaoQiang Li. Face Detection Based on Skin ColorPattern clustering[J]. ShangHai University Journal, 2007,13（5） : 511-515

[11] Rachid Belaroussi,Maurice Milgram.A Comparative Study on Face Detection and Tracking Algorithm[J]. Expert System with Applications, 2012,39(12):7158-7164

[12] AiPing Wang, GuoWei Wan, ZhiQuan Cheng. Incremental Extremely Random Forests Classifier which Support Online Learning[J].Journal of Software, 2011,22（9） : 2059-2074

Discriminative sparse representation and online dictionary learning for target tracking

HuangYue
school of IoT technology
Wuxi institute of technology
Wuxi, China
huangyuewx@163.com

Peng Li
school of IoT Engineering
Jiangnan University
Wuxi, China
pengli@jiangnan.edu.cn

***Abstract*—Traditional sparse representation can not effectively distinguish between target and background. Aiming at these problems, a discriminative sparse representation was proposed, and a discriminative function to the traditional sparse was added for greatly reducing the influence of interference factors. While an online dictionary learning algorithm based on discrimination sparse representation and probabilistic mode was proposed to upgrade target template. It can effectively reduce the impact of the target and the background of the target template. The proposed tracker was empirically compared with state-of-the-art trackers on some challenging video sequences. Both quantitative and qualitative comparisons showed that our proposed tracker was superior and more stable.**

Keywords-sparse representation; discriminative function; dictionary learning; probabilistic mode; target template

I. INTRODUCTION

Target tracking has important applicative value in the intelligent monitoring, human-computer interaction, robot navigation, automatic driving and automatic traffic control. Many scholars have proposed many excellent tracking algorithms, but many interference factors restricted the development of target tracking algorithm. These interference factors include noise, light and shade, rapid movement, sudden movement and so on. Yi Wu [1] details the main methods of tracking algorithm in recent years which including generative and discriminative methods. The generative methods [2-4] represent the object with an appearance model and search for the regions which are the most similar to the object template. Discriminative methods cast tracking as a classification problem [5-6] which considers the tracked object and the background as belonging to two different classes.

Mei proposed the L1 tracker (L1T) [7] for object tracking under the particle filter method based on the sparse represent. L1T describes the tracking target using basis vectors which consist of object templates and trivial templates, and reconstructs each candidate (particle) by a sparse linear combination of them. While object templates correspond to the normal appearance of objects, trivial templates are used to handle noise or occlusion. L1T improves the robustness to occluded object, but sparse representation based trackers perform computationally expensive and prone to drift away from the target in cases of significant changes in appearance. Aiming at the existing problem of L1T, appeared a series of new methods in recent years. A minimal error bounding strategy is introduced [8] to reduce the number of particles, equal to the number of the L1 norm minimizations for solving. A speed up by four to five times is reported in [8]. More recently, accelerated proximal gradient approach was proposed to improve the efficiency of L1T[9]. In [10], the author combines generative trackers and discriminative trackers to give a hybrid approach. The Online Robust Non-negative Dictionary Learning tracker (ONNDL) [11] utilizes dictionary learning to update object template, which made them better adapt to the change of the target.

Dictionary learning (DL) aims to learn from the training samples the space where the given signal could be well represented or coded for processing. One representative DL method for image processing is the KSVD algorithm[12], which learns an over-complete dictionary from a training dataset of natural image patches. Later, Naiyan Wang[13] proposed a probabilistic approach to online dictionary learning, it is formulated with a Laplace error and a Gaussian prior which correspond to an ℓ_1 loss and ℓ_2 regularizer, respectively. It can make reconstructive signal less vulnerable to outliers.

In this paper, we propose a discriminative sparse representation algorithm in the framework of particle filter. IA discriminative function was added in tradition sparse representation which can significantly reduce the probability of target drift and improve the accuracy of target tracking. We also present an online dictionary learning algorithm for updating the object templates. Through iterative calculate discriminative sparse representation and online dictionary proposed in [19] making the updated target template in dictionary can not only robust reconstruct target with interference, but also more accurately distinguish between target and background.

II. PARTICLE FILTERS FOR VISUAL TRACKING

Particle filter algorithm is a good way to solve complex problems in nonlinear and non-gaussian environment, there is no limit to the target state and observation models such as distribution and has been widely used in target tracking. This algorithm including two steps of the forecast and update.

$$p(x_t \mid z_{1:t-1}) = \int p(x_t \mid x_{t-1})\, p(x_{t-1} \mid z_{1:t-1}) dx_{t-1} \quad (1)$$

$$p(x_t \mid z_{1:t}) \propto p(z_t \mid x_t)\, p(x_t \mid z_{1:t-1}) \quad (2)$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.88

Where $p(x_t \mid x_{t-1})$ denotes state transition probability. $p(z_t \mid x_t)$ is the likelihood function of state x_t. The optimal state is obtained by the maximum a posteriori estimation (MAP) over a set of N samples based on sequential MonteCarlo(SMC) method.

III. DISCRIMINATIVE SPARSE REPRESENTATION FOR OBJECT TRACKING

This section is introduced in detail utilizing discriminative sparse representation to reconstruct candidates over target template and backgrounds in dictionary.

In the current frame t, we extracted n particles $Y \in \mathbb{R}^{m \times n}$ according to the above steps. The sparse coefficient of each particle is computed by:

$$J_X = \arg\min_X \{ r(Y,U,X) + \lambda \|X\|_1 \} \tag{3}$$

Where, $U=[U_1\ U_2] \in \mathbb{R}^{m \times r}$, $r = r_1 + r_2$ is composed of target template and background template. $\|X\|_1$ is sparse coefficient constraint, X is sparse coefficient of Y over U and λ is regularization constant. $r(Y,U,X)$ is discriminative function.

The discriminative function guarantee to learn a dictionary can better reconstructing tracking target through the viewpoints of both global and local. From a global perspective, required $Y \approx UX$. From local perspective, it is expected that Y should be well represented by U_1 but bot by U_2. This implies that $\|Y - U_1X_1\|_2^2$ and $\|U_2X_2\|_2^2$ should be small. Thus we define the discriminative term as:

$$r(Y,U,X) = \|Y - UX\|_2^2 + \|Y - U_1X_1\|_2^2 + \|U_2X_2\|_2^2 \tag{4}$$

The objective function is obtained by the above process can be general expressions:

$$\begin{aligned} J_X = \arg\min_X \{ \sum_{i=1}^{n} [& \|y_i - Ux_i\|_2^2 + \|y_i - U_1x_i^1\|_2^2 \\ & + \|U_2x_i^2\|_2^2] + \lambda \|x_i\|_1 \} \end{aligned} \tag{5}$$

Where, $y_i \in Y$, $x_i \in X$, $x_i = [x_i^1; x_i^2]$. Then we can see that all the terms in Eq. (5) except for $\|X\|_1$, are differentiable, and Eq. (5) is strictly convex. The Iterative Projection Method (IPM) in [15] can be employed to solve Eq. (5).

We defined the likelihood value H_c for every candidate

$$H_c = \frac{1}{1 + \exp(-(\xi_2 - \xi_1)/\eta)} \tag{6}$$

Where, $\xi_1 = \|y_i - U_1x_i^1\|_2^2$ is the reconstruction error for y_i with the target template U_1, $\xi_2 = \|y_i - U_2x_i^2\|_2^2$ is the reconstruction error using the background template. The variable η is fixed to be a small constant. So we choose the candidate with maximized H_c to the current target.

IV. ONLINE DICTIONARY LEARNING

After some frames have processed in the video sequence, it is necessary to update the target templates represented to reflect the changes in appearance. To this end, we put forward a kind of online dictionary learning algorithm. We define a matrix $Z=[z_{t-c+1},\cdots,z_{t-i},\cdots,z_t]$ in which each column represents the tracking result of one of the c frames processed before t frame. We formulate it as a robust dictionary learning problem similar to Eq. 7 except that U is now also a variable:

Target and background are certain changes will occur. Background change great, the background of successive frame has no relationship, so we extract the backgrounds U_2 around the new target. However, the change of the target is continuous, we train target template of online by new and old targets. Define the dictionary learning function as:

$$J_{U_1} = \arg\min_{U_1} \|z_t - U_1X_1\|_1^1 \tag{7}$$

Where, z_t is the tracked target of t th, X_1 is sparse coefficient solved by Eq.(5). [19] testified that dictionary through probabilistic mode based on the ℓ_1 loss is robust against external interference. U_1 can be online learning by A_t and B_t. The specific process of online dictionary as follow:

1: Input: z_t; A_{t-1}、B_{t-1}; $U_1 = U_1^{t-1}$; $j = 1,\cdots,m$; m the number of rows of matrix.

2: Starting iteration

3: solved optimal solution of x_1 according to Eq.(5)

$$\omega = \frac{1}{|z_t - U_1x_1| + \sigma}$$

4: *for* $j = 1:m$ *do*

$$A_t^j = A_{t-c}^j / \rho - \frac{\omega^j A_{t-c}^j x_1 x_1^{\mathrm{T}} A_{t-c}^{j\mathrm{T}}}{\rho(\rho + \omega^j x_1^{\mathrm{T}} A_{t-c}^j x_1)}$$

$$B_t^j \leftarrow \rho B_{t-c}^j + \omega^j x_1 z_t^{\mathrm{T}}$$

$$U_1^j = A_t^j B_t^j$$

5: *end for*

6: end of the iteration

7: Output $U_1^t = U_1$, A_t, B_t

Where, ρ is a forgetting factor which gives exponentially less weight to past information. σ is a small constant. When reach the maximum cycles or the difference between adjacent two iterations of U_1 is less than a preset value, end of the iteration

V. EXPERIMENTS

In order to verify the performance of the proposed tracking algorithm in this paper, it compared with several state-of-the-art trackers including the ALSA[4] 、ivt[2]、L1APG[13] 、FCT[7] and LSST[5] tracker on some

challenging video sequences of board sequence, car11 sequence, davidin300 sequence, ThreePastShop2cor sequence, jumping sequence and woman sequence. The parameters of the proposed tracking algorithm are fixed in all experiments. The particle filter uses 600 particles. Weighting parameter is defined as: $\lambda = 0.01$, σ=0.00001, η=0.2。

The center distance err frame-by-frame for each video sequence had been showed in Fig. 1. It indicates that our algorithm has the minimum of the center distance err, and the center distance err in our algorithm has the maximize stability in this frames.

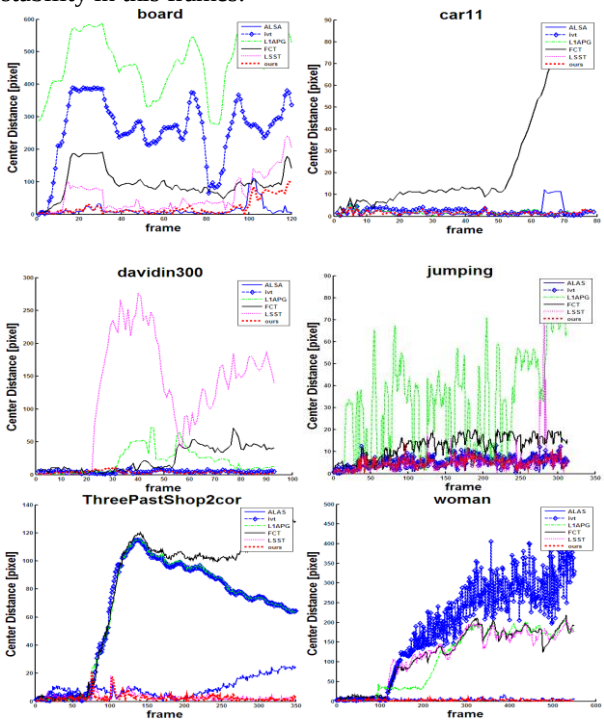


Figure 1 Frame-by-frame comparison of 5 trackers on 6 video sequences in terms of distance score (in pixels).

A. *Quantitative Comparison*

In this paper, two common performance metrics of center distance err and overlap rate are used. As for center distance err, it is the Euclidean distance (in pixels) between the centers of S_{gt} and S_{tr}. S_{gt} and S_{tr} represent bounding box of ground truth and tracker. The overlap rate is defined by overlap percentage which is decided by $\frac{area(S_{gt} \cap S_{tr})}{area(S_{gt} \cup S_{tr})}$. For each frame, a tracker is successful if the overlap rate exceeds 50%. The tracker successful rate is the ratio of the total number of successful frames with the total number of frames.

Table 1 reports the average center location errors in pixels. The value lower, the algorithm accuracy higher. Our algorithm has the highest algorithm accuracy in these challenging videos. Table 2 reports the average success rate, where larger average scores mean that the more accurate results. Our algorithm has the optimum performance in success rate compare with other algorithm in these benchmark videos.

Table 1 central-pixel error (in pixels)

	Our	**ALSA**	**ivt**	**L1APG**	**FCT**	**LSST**
board	20.7	21.1	261.4	459.6	96.3	56.9
car11	1.53	2.23	2.24	1.71	26.5	1.68
davidin300	3.6	3.8	4.7	19.9	23.6	123.4
jumping	4.2	4.8	5.7	33.5	12.6	5.2
ThreePastShop2cor	2.4	9.5	66.2	66.0	80.9	2.9
woman	2.28	4.49	187	110	120	118

Table 2 success rate(%)

	Our	**ALSA**	**ivt**	**L1APG**	**FCT**	**LSST**
board	90	81	5	0	10	15
car11	100	91	98	98	17	100
davidin300	100	100	100	70	90	20
jumping	98	91	90	16	25	93
ThreePastShop2cor	99	70	22	21	22	99
woman	100	99	21	17	20	20

B. *Qualitative Comparison*

With limited space available, a qualitative comparison was showed in Fig. 2 by selected some sequences from complete video sequences randomly.

In the ThreePastShop2cor and woman sequences, the targets were occluded heavily and scale changed. Only our tracker can successfully tracked target in the two sequences in a large scale change with heavy occlusion. It is due to discriminative sparse representation in our algorithm that can distinguish between target and background. In addition, the update scheme of online dictionary learning avoided to choose heavy occlusions scheme which may lead to drift problem.

The board sequence is challenging as the object region undergoes scale variation and out-of-plane rotations. Only ALSA and our tracker had better result, and our tracker owns the most accurate result. The background information was excluded when online dictionary learning, so almost no background information in the dictionary. Therefore, the tracking result will not shift to the background when the board rotates

For the car11 sequence, there is illumination changes and low contrast between the foreground and the background. All trackers tracked well except that FCT tracker. That's because FCT tracker is a discriminative tracker and car11 sequence is low contrast.

About the davidin300 sequence, the target object exist out-of-plane rotations and illumination changes. LSST

tracker drift from target about in 110 frames because illumination changes. FCT and L1APG tracker lost target about in 280 frames because out-of-plane rotations.

In the jumping sequence, the movement is vague and violent. L1APG tracker lost target when the first jumping. However, other trackers tracked the target successfully until the end. Among all algorithms, ours and ALSA get the most accurate result.

Figure 2 Sample tracking results of evaluated algorithms on six image sequences

VI. CONCLUSION

In this paper, a novel moving object tracking algorithm based on discrimination sparse representation is proposed to compute the target template and the particle similarity. The algorithm overcomes tending to drift while tracking target. Moreover, it also can effectively resist the influence of background and heavy occlusion while target tracking. An online dictionary learning algorithm on the basis of discrimination sparse representation is used to update the target template. The biggest advantage is automatically reducing the influence of background change and heavy occlusion while the target template is being updated. So as to achieve more accurate and stable moving target tracking .

REFERENCES

[1] Y. Wu, J. Lim, M-H. Yang, "Online object tracking: A benchmark," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), IEEE Press, June. 2013, pp. 2411 – 2418, doi: 10.1109/CVPR.2013.312.

[2] D. Ross, J. Lim, R. Lin, M-H. Yang, "Incremental learning for robust visual tracking,". International Journal of Computer Vision, vol. 77, May. 2008 pp. 125–141. doi:10.1007/s11263-007-0075-7.

[3] X. Jia, H Lu, M-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), IEEE Press, June. 2012, pp. 1822–1829, doi: 10.1109/CVPR.2012.6247880.

[4] Wang. D, Lu. H, Yang. M. H, "Least soft-threshold squares tracking,"Computer Vision and Pattern Recognition (CVPR), IEEE Press, June. 2013, pp. 2371-2378, doi: 10.1109/CVPR.2013.307.

[5] K. Zhang, L. Zhang, and M. Yang, "Fast Compressive Tracking," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp. 2002-2015, Oct. 2014..

[6] Martin. D, Fahad. S. K, Michael. F, Joost. V. D. W, "Adaptive Color Attributes for Real Time Visual Tracking," Conference on Computer Vision and Pattern Recognition(CVPR), IEEE Press, June. 2014, pp. 1090 – 1097, doi: 10.1109/CVPR.2014.143

[7] X. Mei, H. Ling, "Robust visual tracking using l1 minimization," IEEE International Conference on Computer Vision (ICCV), IEEE Press, Sep. 2009, pp. 1436–1443, doi: 10.1109/ICCV.2009.5459292

[8] X. Mei, H. Ling, Y. Wu, E. Blasch, L. Bai, "Minimum error bounded efficient L1 tracker with occlusion detection," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), IEEE Press, June. 2011, pp. 2661–2675, doi: 10.1109/TIP.2013.2255301.

[9] C. Bao, Y. Wu, H. Ling, H. Ji, "Real time robust L1 tracker using accelerated proximal gradient approach," Conference on Computer Vision and Pattern Recognition(CVPR), IEEE Press, June. 2012, pp. 1830–1837, doi: 10.1109/CVPR.2012.6247881.

[10] W. Zhong, H. Lu, M-H. Yang, "Robust object tracking via sparsity-based collaborative model,"Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), IEEE Press, June. 2012, pp. 2356–2368, doi: 10.1109/CVPR.2012.6247882.

[11] Wang. N, Wang. J, Yeung. D-Y, "Online robust non-Negative dictionary learning for visual tracking," IEEE International Conference on Computer Vision (ICCV), IEEE Press, Dec. 2013, pp. 657–664, doi:10.1109/ICCV.2013.87.

[12] Aharon. M, Elad. M, Bruckstein. A, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,"[J]. IEEE Transactions on Signal Processing, vol. 54, pp. 4311-4322, October 2006.

[13] Wang. N, Yao. T, Wang. J, Dit-Yan. Y, "A probabilistic approach to robust matrix factorization,"Computer Vision–ECCV 2012. Oct. 2012, pp. 126-139, doi: 10.1007/978-3-642-33786-4_10.

[14] [14] M. Yang, L. Zhang, "Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary,"11th European Conference on Computer Vision (ECCV), Sep. 2010, pp. 5-11 doi: 10.1007/978-3-642-15567-3_33.

A Secure Blind Watermarking Scheme Based on Embedding Function Matrix

Wang Xiao
School of Information Technology Engineering
Tianjin University of Technology and Education
Tianjin 300222,China
syowang@vip.163.com

Liu Shuo
School of Information Technology Engineering
Tianjin University of Technology and Education
Tianjin 300222,China
446809295@qq.com

***Abstract*—In this paper, a secure blind watermarking scheme is proposed. In the scheme, an embedding function matrix(EFM) is generated based on a serial number which is unique to each owner or owner group, then the cellular automata sys-tem embeds the scrambled watermark into the 3-level DWT coefficients of the original image based on the EFM. In the procedure of extracting, the original image is not needed, however even if the algorithm is known, without the correct function matrix, watermark can not be extracted. This shows that the pro-posed scheme can enhance the security of watermark. Experimental results also show that the proposed scheme is robust to geometrical attacks and common image processing attacks**

Keywords- watermark; embedding function matrix; cellular automata;

1 Introduction

In the past one decade, digital image watermark has been an important branch of information security, which is an effective technique for copyright protection, authentication etc. Digital image watermark technique requires many factors such as capacity, invisibility, robustness and security. In this paper, we will emphasis on the factor of security. In the past few years, the security of watermark has been focused mainly on scrambling algorithms and the domains for embedding, many schemes and algorithms have been proposed[1-5]. However, according to the Kerckhoffs's principle, the security of watermark should depend on the key but not on the algorithm. In this paper, a secure blind watermarking scheme based on embedding function matrix is proposed. Firstly, 3-level DWT is performed to the original image to obtain the DWT coefficients, meanwhile, an embedding function matrix(EFM) is generated based on a serial number which is unique to each owner or owner group, then the cellular automata system embeds the scrambled watermark into the 3-level DWT coefficients based on the EFM. In the procedure of extracting, original image is not needed, however even if the algorithm is known, without the correct function matrix, watermark can not be extracted. From the experimental results, we can see that the proposed scheme can enhance watermark security and is robust to geometrical attacks and common image processing attacks.
The paper is organized as follows: section 2 presents the description of the proposed scheme, section 3 presents the experimental results, and section 4 is the conclusions.

2 Watermark Embedding

In this work, we use a gray image as original image and a binary image as watermark. Assume the original image is O, the binary watermark is W, $W=(w_{ij})$, $w_{ij}\in\{0,1\}$, (i,j) denotes a pair of coordinates, the serial number (SN) is designed by the issuer of image, which is a 64bit hexadecimal number sequence.
The embedding steps are shown as follows:

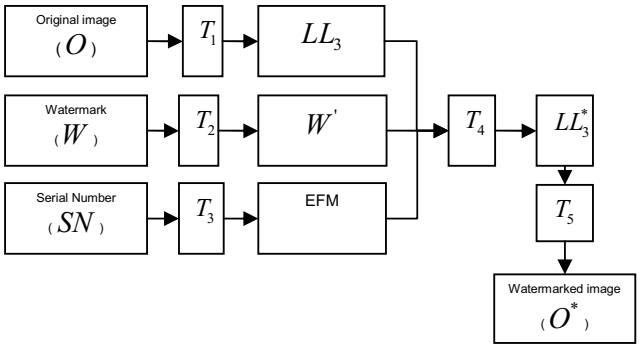


Fig. 1. Watermark Embedding

2.1 DWT

T_1 denotes 3-level DWT decomposition, which is performed to O to obtain the 3-level low-frequency coefficient set LL_3 as the watermark carrier. In this paper, the wavelet base is Haar wavelet. T_5 denotes 3-level inverse DWT transformation.

2.2 Arnold Transformation

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.89

T_2 denotes 2D Arnold transformation, which is performed to the original watermark, 2D Arnold transformation is the first level encryption by changing the distribution of W,

the algorithm of T_2 is defined as follows:

$$\begin{cases} X' = (X+Y)\%N+1 \\ Y' = (X+2Y)\%N+1 \end{cases} \quad (1)$$

Where (X,Y) are the pixel of the original watermark and (X',Y') are the pixel scrambled image. After T_2 ,the original watermark W is converted to W'.

2.3 Embedding Function Matrix

The purpose of T_3 is to generate a two dimensional binary matrix as the embedding function matrix. The function matrix is decided by the serial number. Theoretically, there are many optional method for T_3 ,and each method is also as an encrypting key. In this work, we used a simple method, the steps are as follows:

Step 1. convert the 64bit hexadecimal serial number into 256bit binary serial number(SN).

Step 2. define a 64×64 function matrix(M), which is divided into 256 sub matrixes of 4×4 ,for each sub matrix

$$M_{(4\times i-3:4\times i,4\times j-3:4\times j)} = \begin{cases} 1 & if \quad SN_{(16\times i+j)} = 1 \\ 0 & if \quad SN_{(16\times i+j)} = 0 \end{cases} \quad (2)$$

Where $SN_{(k)}$ denotes the value in SN at position k. Theoretically, the matrix space is 2^{256} . Figure 2 shows some function matrixes generated based on different SN as examples.

Fig. 2. examples of embedding function matrixes

2.4 Cellular Automata Transform

T_4 denotes cellular automata transform which embeds the watermark into the carrier based on the EFM. In this work, we use Moore neighborhood model to define the cellular automata system. The Moore neighborhood comprises the 8 pixels surrounding a central pixel on a 2-dimensional square lattice. Moore neighborhood model is defined as follows:

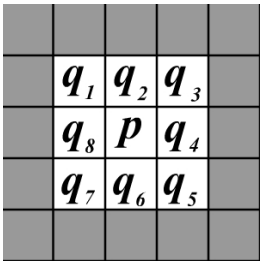


Fig. 3. Moore neighborhood model

The embedding method is defined as follows:

$$p^*_{(i,j)} = \begin{cases} \bar{q}_{(i,j)} + 0.5w_{(i,j)} & if\ f_{(i,j)} = 0 \\ \bar{q}_{(i,j)} - 0.5w_{(i,j)} & if\ f_{(i,j)} = 1 \end{cases} \quad (3)$$

$\bar{q}_{(i,j)}$ denotes the average of gray scale in Moore neighborhood ($\sum_1^9 q_i$)

$f_{(i,j)}$ denotes the embedding function at (i,j).

$w_{(i,j)}$ denotes the value of watermark at (i,j).

$p^*_{(i,j)}$ denotes the modified pixel from the original pixel $p_{(i,j)}$.

When all the pixels are modified, we obtain the watermarked low-frequency coefficient set (LL_3^*).

When the steps $T_1 \sim T_5$ are performed, we obtain the watermarked image(O^*).

3 Watermark Extraction

When ownership issue occurs, the arbitration agency can perform watermark extraction to judge the ownership. The extracting steps are shown as follows:

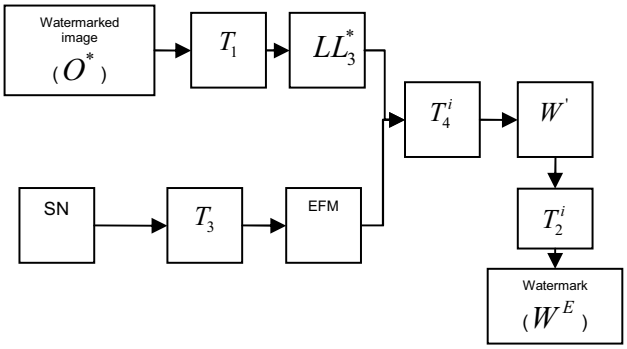


Fig. 4. Watermark extraction

T_1, T_3, O^*, LL_3^* EFM, W' and SN has the same definition as in section 2. T_2^i denotes the inverse Arnold transformation. W^E denotes the extracted watermark.

T_4^i is described as follows:

$$w_{(i,j)} = \begin{cases} \dfrac{p^*_{(i,j)} - \bar{q}_{(i,j)}}{0.5} & if \ \ f_{(i,j)} = 0 \\ \dfrac{\bar{q}_{(i,j)} - p^*_{(i,j)}}{0.5} & if \ \ f_{(i,j)} = 1 \end{cases} \tag{4}$$

$p^*_{(i,j)}, \bar{q}_{(i,j)}, f_{(i,j)}, w_{(i,j)}$ has the same definition as in T_4.

4 Experimental Results

In the experiment, the original image is the popular Lena image of size 512×512, the watermark is of size 64×64.

Figure 5 presents the original Lena image(O) ,the watermark image(W) the watermarked Lena image(O^*), and the extracted watermark (W^E) under no attacks. The embedding function matrix and the extracting function matrix are the same function matrix, which is generated based on a random SN. The PSNR of O^* is 37.2122dB, which shows the proposed scheme has good invisibility.


(a1)Original image


(b1)Watermarked image


(a2)Original watermark


(b2)Extracted watermark

Fig. 5. Embedding and Extracting with the same function matrix

Figure 6 presents the original Lena image(O) ,the watermark image(W) the watermarked Lena image(O^*), and the extracted watermark (W^E).The embedding function matrix and the extracting function matrix are different function matrixes. The watermark can not be extracted, which shows the proposed scheme has good security.


(a1)Original image


(b1)Watermarked image


(a2)Original watermark


(b2)Extracted watermark


(a3) Embedding function matrix


(b3) Extracting function matrix

Fig. 6. Embedding and Extracting with different function matrixes

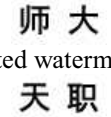
Figure 7 presents experimental results under attacks, which shows the proposed scheme is robust to geometrical attacks and common image processing attacks


(a1)the right part is cropped


(b1)the underside part is cropped


(c1)the central part is cropped


(a2)Extracted watermark from a1


(b2)Extracted watermark from b1

(c2)Extracted watermark from c1

(d1)0.2% Gaussian attack

(e1)0.2% salt and pepper noise attack

(d2)Extracted watermark from d1

(e2)Extracted watermark from e1

Fig. 7. Experimental results under attacks

5 Conclusions

The paper proposed a secure blind watermarking scheme based on embedding function matrix. The cellular automata system embeds the scrambled watermark into the 3-level DWT coefficients with the EFM. and without the correct function matrix, watermark can not be extracted. This shows that the proposed scheme can enhance the security of watermark. The experimental results show that the proposed scheme is robust to geometrical attacks and common image processing attacks.

References

[1] IJCox. A Secure, Robust Watermark for Multimedia[J]. Proc Information Hiding First International Workshop, 1996.

[2] Shiba R, Kang S, Aoki Y. An image watermarking technique using cellular automata transform[C]// TENCON 2004. 2004 IEEE Region 10 Conference. IEEE, 2004:303 - 306 Vol. 1.

[3] Langelaar G C, Setyawan I, Lagendijk R L. Watermarking digital image and video data. A state-of-the-art overview[J]. Signal Processing Magazine IEEE, 2000, 17(5):20 - 46.

[4] Zhen-Yu G U, Yuan Y, Zhang G F. Blind Watermark Scheme Based on Masking Effect and Dither Modulation[J]. Packaging Engineering, 2013.

[5] Agarwal H, Raman B, Venkat I. Blind reliable invisible watermarking method in wavelet domain for face image watermark[J]. Multimedia Tools & Applications, 2014.

[6] Podilchuk C I, Delp E J. Digital watermarking: algorithms and applications[J]. IEEE Signal Processing Magazine, 2001, 18(4):33 - 46.

[7] Wang Y, Doherty J F, Van Dyck R E. A Wavelet-based Watermarking Algorithm for Ownership Verification of Digital Images[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2002, 11(2):77--88.

[8] Kang X, Huang J, Shi Y Q, et al. A DWT-DFT composite watermarking scheme robust to both affine transform and JPEG compression[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2003, 13(8):776--786.

[9] Agarwal H, Raman B, Venkat I. Blind reliable invisible watermarking method in wavelet domain for face image watermark[J]. Multimedia Tools & Applications, 2014.

Image segmentation method combines MPM / MAP algorithm and geometric division

LINGHU Yong-Fang
Guizhou Colloge of Finance and Economics
GuiYang ,China
lhyongfang@aliyun.com

Shu Heng
Guizhou Normal University
GuiYang.China
Shuheng_mast@aliyun.com

***Abstract*—A novel image segmentation algorithm based on a Bayesian framework is studied in this paper.We presents a new region and statistics based approach， which combines Voronoi tessellation technique and Maximum a posterior / Maximi-zation of the posterior marginal (MAP /MPM) algorithm. The image domain is partitioned into a group of sub-regions by Voronoi tessellation， each of which is a component of homogeneous regions. And the image is modeled on the supposition that the intensities of pixels in each homogenous region satisfy an identical and independent gamma distribution.The initial segmentation is applied to obtain number of the initial motions and the corresponding initial parameters of the image model.Then the parameters are updated by using the given parameter estimation method. A fast estimation procedure for the posterior marginals is added to the MAP algorithm.The experiment results show that the proposed algorithm here is effective.**

Keywords- Voronoi tessellation; Maximi-zation of the posterior marginals; MAP algorithm； Image segmentation

I. INTRODUCTION

In the research and analysis of the image, people often to the specific, unique in the images with regional interest, image segmentation is refers to the technology and process of these areas are extracted. Because of image segmentation as a frontier subject is full of challenges, in recent years, attracted many scholars engaged in the research of this field. Image segmentation technology in aerospace, biomedical engineering, industrial detection, robot vision, public security, judicial, military guidance, culture, art, geography, surveying and mapping, and other fields has been widely attention, and a great pioneering achievements, make it become a dramatic and broad prospects of the new discipline.

Mainly including based on region segmentation, based on the boundary of the segmentation, the segmentation based on clustering and segmentation based on statistical model, among them, the image segmentation method based on statistical model is most noteworthy [1]. The segmentation algorithm involving major problems include: to establish a statistical model, model parameter estimation of the image as well as to establish the best integral optimization criterion. In the process of image statistical model is set up the introduction of spatial information and spatial correlation can improve the accuracy and reliability of the algorithm, one of the most representative is Markov random field(MRF) model [2].

In order to solve the existing model parameter estimation and obtain the best segmentation rule of optimization problems, this paper applies MAP/MPM to statistics in the image segmentation method based on region. First based on Voronoi division, image model is set up; And then within the framework of bayesian theory to establish image segmentation model; Finally combining with the MAP /MPM algorithm for image segmentation and model parameter estimation.

II. IMAGE MODEL

In order to realize the image segmentation, this paper generalized MRF image based on Voronoi partition model is set up.

A. Voronoi tessellation

Assuming that involves uniform distribution in image domain D, and the distribution of each involves are independent of each other.[3] Therefore, involves the collection of prior probability density function is:

$$p(G|m) = \prod_{j=1}^{m} p(u_j, v_j) = \prod_{j=1}^{m} \frac{1}{|D|} = |D|^{-m} \qquad (1)$$

Among them, for the area of the image domain D. Such as hypothesis generating points m meet averages poisson distribution, then the probability density function is,

$$p(m) = \frac{\lambda^m}{m!}\exp(-\lambda) \qquad (2)$$

With the traditional pixel as basic processing unit of image segmentation methods, the proposed approach with Voronoi polygon of the basic processing unit, the assumption in the same Voronoi polygon within the pixels belonging to the same target class. Thus, for each Voronoi polygon P_j distribution a label variable $L_j \in \{1,\cdots,k\}$, to characterize the membership of the target class, where k to be the goal of image segmentation total, and each target class of geometric re gion is by a group of Voronoi polygon fitting.

It is obvious that a random label field is formed for the label set $L = \{L_j : j = 1,\cdots,m\}$ of all Voronoi polygons,

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.90

and each of the random fields of the random field corresponds to a segmentation result of R. Similarly, the marking places may realize the constitution of space recorded as Ω_L In order to express domain Voronoi polygon label correlations, this paper will define the rules on the lattice of MRF model is extended to the Voronoi graph partition. It is assumed that the label field L probability density function has the following form:

$$p(L|G,m)=\frac{1}{A}\exp\left(-\sum_{(L_j,L_{j'})\in NP}\beta t(L_j,L_{j'})-\sum_{L_j\in L}\gamma_{L_j}\right) \quad (3)$$

In the formula, NP is a collection of all the neighboring Voronoi polygons in the given Voronoi partition, and any two Voronoi polygons are mutually domain and only if they have a common boundary. A is the normalized constant, and is equal to all the possible labeled values of the formula (3).

B. Neighborhood model

On the basis of Voronoi division，Neighborhood model is based on the same subdomain domain was realized by modeling the correlation between pixels. The correlation can be made by subdomain P_j pixels (x_i,y_i) intensity of Z_i conditional probability density function $p(Z_i|Z_{Ni},\theta_{Lj})$, Among them $Z_{N_i}=\{Z_{i'};(x_{i'},y_{i'})\in N_i\bigcap p_j\}$, N_i pixels in pixel (xi,yi) field collection, why label L associated with a set of parameters. For the sake of simplicity, field to use pixel for binary field of Gaussian distribution to establish the correlation between pixels. And assumptions: 1) the subdomain P_j label $L_j=l$, among them, the intensity of all pixels with the same mean and variance, respectively, for μl and σl;2) pixels to the covariance matrix of covariance related to field of pixels in the direction of δ .As shown in figure1 (a)X said don't consider the points, 0 indicates its adjacent points.Second order neighborhood system (8 neighborhood systems) in every point within eight adjacent points, as shown in figure1 (b) . Figure1 (c) the number in the n = 1, 2,... Said, 5 in the n order neighborhood the outermost layer of the adjacent points in the system.

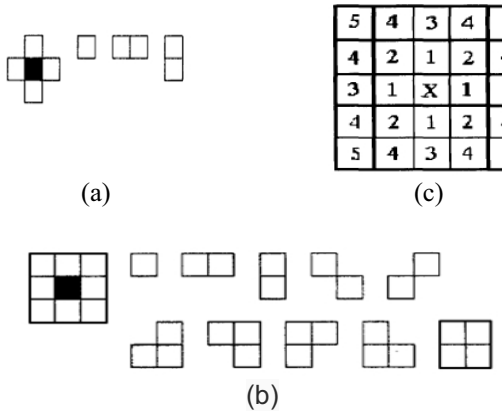


Figure 1 Neighborhood relation and potential clique type graph

S for the set of nodes, N based on the neighborhood relationship of the connections between nodes, it (S, N) constitute the usual sense of the figure. Figure (S, N) of a group of potential is defined as a S a subset of the set of all meet the adjacent relationship between different nodes (except the single point potential regiment). Figure 3.1 shows the potential of first and second order neighborhood system type. As shown in figure1(a) a single point of horizontal and vertical direction, the two potential group for all types of first-order neighborhood systems, second order neighborhood system in addition to this, and diagonal situation, three power group and four potential regiment (figure1 (b)). It can be seen that as the growth of the order number, the number of potential regiment followed a sharp rise.

C. Markov random field model

The c is on a two-dimensional image A a clique. That c is a subset of A. And c is formed by a single point or c all the points are adjacent to each other. $C_s=\{c|s\in c\}$ is a collection of cliques.

For each image point marked b, the probability distribution can be described using the following formula

$$P(b)=\frac{1}{Z_b}\exp\left[\beta\sum_{\langle r,s\rangle}V(b(r),b(s))\right]$$

where $P(b)$ is a prior indicator distribution, Z_b is the normalizeing constant, β is a positive parameters that controls the granularity of the regions,the sum is taken over all nearest-neighbor pairs of sites in L and the Ising potential V is:

$$V(b(r),b(s))=\begin{cases}-1 & if\ \sum_{k=1}^{K}b_k(r)b_k(s)=1,\\ 1 & otherwise\end{cases} \quad (4)$$

(Note that to all the $r\in L, b_k(r)\in\{0,1\}$ and $\sum_{k=1}^{K}b_k(r)=1$, so that $V(b(r),b(s))=-1$ if r and s belong to the same segment.)

Calculated using the Bayesian posterior distribution rules, and its form is:

$$P(b,\theta|g)=\frac{P(g,b|\theta)P(\theta)}{P(g)}=\frac{1}{Z}\exp[-U(b,\theta)]$$

Where Z is a standardized constant, and：

$$U(b,\theta)=-\sum_{r\in L}\sum_{k=1}^{K}b_k(r)\log l_k(r)+\beta\sum_{\langle r,s\rangle}V(b(r),b(s))+\log P(\theta) \quad (5)$$

assumes that $l_k(r)$ an observation noise model, in line with the gaussiandistribution.

$$l_k(r)=\Pr(g(r)|\theta,b_k(r)=1)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left[-\frac{D(g,\Phi(r;\theta_k))}{2\sigma_k^2}\right] \quad (6).$$

III. Posterior edge of the largest combination of probability and the MAP method

Here we discuss an algorithm based on Bayesian estimation theory, based on [6] work, and suitable for simultaneous calculation of the indicator function and the parameter vector estimates (b and θ). In this method, we consider b and θ is a random vector, its best estimate is that by minimizing an appropriate posterior distribution on the expected loss function. It allows the use of rapid methods to estimate the posterior edge.

We will take the loss function is:

$$C(\hat{b},\hat{\theta},b,\theta) = 1 - \delta(\theta - \hat{\theta}) + \frac{1}{|L|}\sum_{r\in L}\left[1 - \delta(b(r) - \hat{b}(r))\right] \quad (7)$$

Here, if their argument is the 0 vector, then the δ function equal to 1, otherwise equal to 0; x expressed in the L on the number of pixels. The first requirement $\hat{\theta}$ of the estimated parameter vector estimate, the average is an unbiased estimate, while the second request by the division of the expected average error minimization to estimate the direction of the support function of the domain $\hat{b}$. Therefore, the optimal estimation is: $(\hat{b}^*,\hat{\theta}^*) = \arg\min_{\hat{b},\hat{\theta}} Q(\hat{b},\hat{\theta}) = E[C(\hat{b},\hat{\theta},b,\theta)]$

To minimize Q we propose a 2-step procedure in which $Q(\hat{b},\hat{\theta})$ is minimized with respect to $\hat{b}$ for a given $\hat{\theta}$ in a first step,and then minimized with respect to $\hat{\theta}$,keeping the optimal $\hat{b}$ fixed,in the second step.To derive the implementation of the first step,we make the following considerations:suppose that $\hat{\theta} = \bar{\theta}$ is given.The optimal estimator for $\hat{b}$ is found by minimizing the expected value of the second term of Eq.(4):

$$\begin{aligned}
&\frac{1}{|L|}\sum_{b}\sum_{r\in L}(1 - \delta(b(r) - \hat{b}(r)))]\,P(b,\bar{\theta} \mid g) \\
&= 1 - \frac{1}{|L|}\sum_{r\in L}\sum_{b:b(r)=\hat{b}(r)} P(b,\bar{\theta} \mid g) \\
&= 1 - \frac{1}{|L|}\sum_{r\in L}\sum_{k=1}^{K}\sum_{b:b(r)=\hat{b}(r)} P(b,\bar{\theta} \mid g)\hat{b}_k(r) \\
&= 1 - \frac{1}{|L|}\sum_{r\in L}\sum_{k=1}^{K}\pi_k(r)\hat{b}_k(r)
\end{aligned} \quad (8)$$

Where $\pi_k(r) = \sum_{b:b_k(r)=1} P(b,\bar{\theta} \mid g)$ (9)

is the posterior marginal probability for the support region k at pixel r.Expression(6) is minimized by setting $\hat{b} = \bar{b}$,with

$$\bar{b}_k(r) = \begin{cases} 1 & if\ \pi_k(r) > \pi_{k'}(r)\ for\ k' \neq k, \\ 0 & otherwise \end{cases} \quad (10)$$

This estimator is called the Maximizer of the Posterior Marginals or MPM estimator[7]for b given $\bar{\theta}$.

The problem is that the edge of the direct calculation of $\{\pi(r)\}$ is not practical, because in equation (9) the sum of the right side too much. Thus, we propose a useful approximation, will be described below.

A.Posterior edge of the fast estimation procedures

Predecessors have put forward many methods to estimate the posterior edge of discrete MRF: Markov chain rule [8] the stochastic method and the certainty based on the mean field approximation method [9], the edge is estimated that large-scale joint solution Li obtained nonlinear equations. However, these method is very expensive. Here, we use [10] of the proposed method, only isolated solutions of linear equations, allows for quick access high-quality estimation, leads to the discrete equations can be very effective over mysterious transformation solved. In general, the MPM-step application to GMMF estimation method based on the following steps:

1. Compute the normalized likelihood field $\{l_k(r), r\in L, k=1,\ldots,k\}$ (which isdirect given by the noise model,Eq.(6)).
2. Posterior edge of the calculation of best estimate $\{p_k(r) \approx \pi_k(r),\ r \in L, k = 1,\ldots,k\}$,

Minimized by the K function Ek, k = 1, ..., K (the use of DCT):

$$E_k(p) = \sum_{r\in L}(p_k(r) - l_k(r))^2 + \lambda\sum_{\langle r,s\rangle}(p_k(r) - p_k(s))^2 \quad (11)$$

3. Compute the MPM estimator for the original field b as:

$$\bar{b}_k(r) = \begin{cases} 1 & if\ p_k(r) > p_{k'}(r)\ for\ all\ k' \neq k, \\ 0 & otherwise \end{cases} \quad (12)$$

B.Complete algorithm

Fixed to b on the s to minimize the Q, just consider (6) The first type of expectation: $\int[1 - \delta(\theta - \hat{\theta})]dP(\bar{b},\theta \mid g) = 1 - P(\bar{b},\hat{\theta} \mid g)$,

Therefore θ been passed on the smallest of $U(\bar{b},\theta)$ (equation (5)) to find the optimal θ estimate (maximum a posteriori probability or MAP).

Complete algorithm is as follows:

1. Vector calculation of the parameters of the initial estimated value of $\theta^{(0)}$, and set t = 0;
2. Implementation (3-5) step-by-step until convergence;
3. MPM Step：

3.1. Calculation of posterior probability estimates $p_k(r),\ r \in L,\ k = 1,\ldots,K$

3.2. By equation (9) calculated $\bar{b}_k(r)$

4．MAP-step: Calculate $\theta^{(t+1)} = \arg\min_{\theta} U(\bar{b},\theta)$ (minimum of equation (6));

5．Set t:=t+1。

Convergence of this algorithm is because there is a lower bound of Q (it is always non-negative), and after a complete iteration does not increase. It may converge to a limited circle, in fact, it converges to a fixed point. $\{\pi_k(r)\}$ on the posterior edge of the estimation error, this algorithm better than the EM

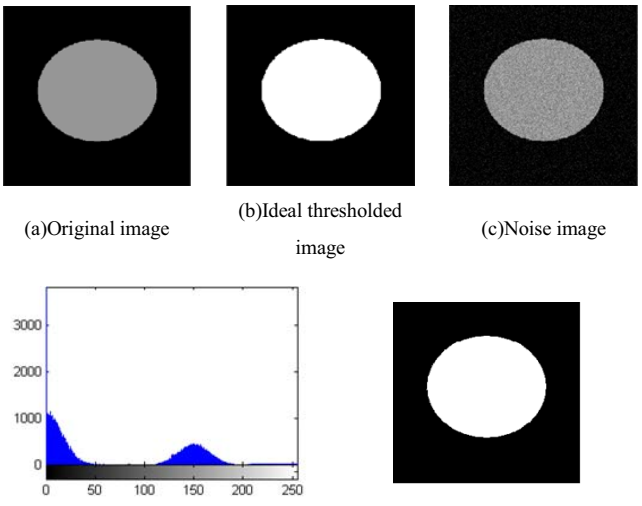
algorithm robustness, because in each model of the distribution of $\{\pi_k(r)\}$, only the location of an important variable.

IV. experiments and their results

We do two types of experiments: First, the use of synthetic image sequences; Second, the real standard sequence. Initial marking field is obtained based on the observed field can be estimated by maximum likelihood method or the method based on histogram, you can simply set a threshold to get the initial label field.

A. The synthetic image experiment

Figure2 (a) is a composite image of the original pixel, in its background region (the grey value of 50) has a circular object (the gray value of 150), the segmentation of the ideal as shown in figure2(b). Figure 2(c) in figure 2 (a) with gaussian noise, the histogram is shown in figure 2(d). Obviously, figure 2 (c) is a synthetic image in the composition of two types of pixel. Figure2(d) is the histogram of the noise image. The final segmentation results are shown in Figure2 (e).



(a)Original image (b)Ideal thresholded image (c)Noise image

(d)Histogram of the noise image (e)segmentation result

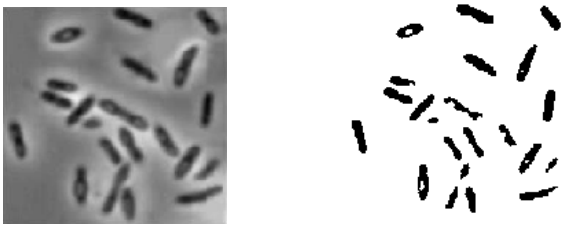
Figure 2 Synthetic Image Sample

B. The real image experiments

In addition to the synthetic images, we also chose 2 real image to verify the experiment. The 2 images were moon.tif and bacteria. Tif.

(a)Moon original image (b)Moon segmentation results

Figure3 Moon Image Sample



(c) Bacteria original image (d)Bacteria segmentation rsults

Figure4 Bacterial Image Sample

V. Conclusion

Image segmentation is a key in the process of image processing and image analysis steps. In this paper, combining with Voronoi partitioning technology and EM/MPM algorithm, realize the image segmentation.The method using the Voronoi technology will image domain is divided into sub regions, and to set up considering neighborhood relations, regional and global pixel of image model. Then model using MPM/MAP algorithm to iteration.. Through experiments, we proposed a priori MRF models.edge based on the bayesian estimation theory of image segmentation algorithm is effective. In the next work, based on the combination of image features, more (such as color, texture, etc.), and to improve the segmentation accuracy and efficiency.

REFERENCES

[1] *MA Long, WANG Lu-Ping, SHEN Zhen-Kang. A slow-moving-object segmentation technology based on MRF un-der complex background [J]. Signal Processing, 2010, 26(6) : 911-916.*

[2] *Comer M, Bouman C A, De Graef M, Simmons J P. Bayesian methods for image segmentation [J]. Journal ofthe Minerals, Metals and Materials Society, 2011, 63(7) : 55-57.*

[3] *Li Y, Li J. Segmentation of SAR intensity imagery with aVoronoi tessellation , Bayesian inference , and reversiblejump MCMC algorithm [J]. IEEE Transactions on Geo-science and Remote Sensing, 2010, 48(4) : 1872-1881*

[4] *Bors Adrian G, Pitas Prediction tracking and Ioannis. It's moving in sequence of dura [J]. IEEE from Processing, 1997,89 (dura. 8) : 1441-1445*

[5] *Felix Calderon, J.L. Marroquin,Salvador Botello,B.C.Vemuri, The MPM-MAP algorithm for motion segmentation,in:Computer Vision and Image Understanding 95(2004)165-183.*

[6] *Felix Calderon, J.L. Marroquin,Salvador Botello,B.C.Vemuri, The MPM-MAP algorithm for motion segmentation,in:Computer Vision and Image Understanding 95(2004)165-183.*

[7] *Li S Z. Markov Random Field Modeling in Computer Vision[M].Tokyo: Springer-Verlag,1995.*

[8] *Y.Weiss, Smoothness in layers:motion segmentation using nonparametric mixture estimation,Proc.IEEE Conf. Comput. Vision Pattern Recogn. (1997) 520-527.*

[9] *J.Marroquin,S. Mitter, T. Poggio, Probabilistic solution of ill-posed problems in computational vision, J.Am.Stat.Assoc. 82(1987.76-89.*

[10] *Yongyue Zhang,Michael Brady,and Stephen Smith.Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expection-Maximization Algorithm,IEEE Transactions on Medical Imaging,2001, 20(1),45-57.*

License Plate Location Based on Quantum Particle Swarm Optimization

chen yuping
Wuxi Traffic Branch
JIANGSU Union Technical Institute
Wuxi, China
Pingcy118@126.com

***Abstract*—This paper presents a quantum particle swarm optimization solutions for the problem of license plate location, the first image of the vehicle local area adaptive illumination compensation image preprocessing, improving the quality of the picture. Then use the YCbCr color model signals a band compression to achieve illumination compensation. Based on this paper proposes a quantum particle swarm optimization algorithm, using less number of QPSO algorithm parameters, randomness is strong, and can cover all the solution space, to ensure global convergence of the algorithm. This paper presents the results of three pieces of plate positioning images prove local feature enhancement algorithm QPSO algorithm can perform feature extraction and stable positioning works well.**

key words- Quantum particle swarm optimization ; Positioning plates

I. INTRODUCTION

License plate positioned in computer vision research has been an important task, however, because the actual image acquisition typically has complex backgrounds, plus the acquisition of image quality vary widely, from an objective image of the target area increases the difficulty of targeting. Intelligent Transportation Systems ITS (Intelligent Transportation System) will be effectively used as the entire transport management system of information technology, data communications transmission technology, electronic control technology and computer processing technology, often require a vehicle license from real shot images were extracted identification of the vehicle in order to complete the authentication. In order to establish a large-scale, full-functioning real time, accurate and efficient transport integrated management system. How to solve the ITS vehicle license positioning is modern intelligent transportation systems in a typical positioning of the target area.

Current common vehicle license plate location algorithm, there are four categories: 1) detect fuzzy mathematical theory to get the license plate location information based on license plate image set straight borders [1]. 2) Edge detection and location identification to the Hough transform-based to get the license plate location [2,7]. 3) license plate based on genetic and neural network location [3-4]. 4) Based on the horizontal direction texture features [5].

However, the existing license plate location algorithm there are two problems: First, complex background interference, which depends on an effective feature extraction, the second is really making the image quality problems that affect the weather, lighting and other conditions often make the image of the license plate area with respect to the full and have a greater degradation difficult to identify.

This paper presents an optimization algorithm of license plate location algorithm based on quantum particle swarm, to a certain extent, a better solution to these two issues, the positioning of adaptability is greatly improved. By local image enhancement processing to overcome the image quality really making changes in the larger issue, so as to obtain the desired plate characterization, combined with quantum particle swarm optimization efficient and fast search for the most characteristic feature of the plate to meet the full range of regional position in order to accurately locate the plate.

II. FEATURES GET

A. *Image preprocessing*

Digital image processing, which is a process of human visual information processing, plays a more and more important role in modern life. Digital image quality has a direct impact on subsequent processing results, so it is necessary to pre process the digital image, in order to achieve the purpose of improving the image quality and ensuring the processing results. Digital image processing is widely used in many fields such as biomedicine, materials, remote sensing, communication, traffic management, military reconnaissance, document processing and industrial automation. Digital image preprocessing is a part of digital image processing. It plays an important role in the field of image processing. The main purpose of the image preprocessing is to eliminate unrelated to the image information, the recovery of the real and useful information and information related to maximize detection and simplify the data, so as to improve the feature extraction and image segmentation, matching and recognition reliability enhancement and so on. The image processing level lags far behind in the world advanced level, the demand of the development of the technology is imminent.

Image Deviation of automatic number plate recognition systems in practical applications, capture images to be affected by many external conditions, such as changes in the

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.91

natural light of day and night, the change of seasons, body height, the viewing angle is different among other factors, compared with the acquired Great, but there are a variety of noise, and therefore must be on the plate image pre-processing to improve image quality. In response to these license plate image, the paper uses an adaptive illumination compensation method for image pre-processing, experiments show that illumination compensation effectively increases the dynamic range of the image colors, enhanced contrast, improved picture quality, for subsequent create conditions for pre-processing operations.

B. local image processing

Weather lighting conditions is also very easy to make the target area at a full background characteristics is not obvious, so a local treatment is an effective way to strengthen the image features can be carried out for the license plate for the target area. YCbCr or Y'CbCr will be written: YCBCR or Y'CBCR, it is a kind of color space, usually used in the image processing, or digital photography system. Y'for the color of the brightness (luma) components, and CR and CB is the concentration of blue and red shift in the amount of components. Y 'and Y is different, and Y is the so-called lumens (luminance), said light concentration and nonlinear using gamma correction (gammacorrection) encoding process.

YCbCr in which Y refers to the luminance component, Cb refers to the blue color component, while the Cr refers to the red color component. The human eye is more sensitive to the Y component of the video, so it can reduce the quality of the image by using the sub sampling to reduce the color component. The main sub sampling formats are 4:2:0 YCbCr, 4:2:2 YCbCr and 4:4:4 YCbCr.4:2:0 said every four pixels have 4 luminance and two chrominance components (YYYYCbCr), sampling only the odd numbered scan lines, portable video device (MPEG-4) and video conferencing (H.263) is the most commonly used format; 4:2:2 said every four pixels 4 luma component, four chroma component (YYYYCbCrCbCr) is most commonly used format for DVD, digital TV, HDTV and other consumer video equipment; 4:4:4 said a full pixel matrix (YYYYCbCrCbCrCbCrCbCr), for high quality video applications, studio and professional video products.

YCbCr model takes full advantage of the human eye is sensitive to luminance signal and chrominance signals are relatively insensitive to this feature, the color difference signal is a band compression, you can retain the original color content on the basis on the use of gray-enhanced process enhanced wherein Y luminance component, so as to achieve the purpose of illumination compensation. So, here we first RGB color space using equation (1) is converted into YCbCr, after finished illumination compensation recycling equation (2) to change back to the RGB model: RGB into YCbCr formula is formula(1)

$$\begin{cases} Y = (77/256)R + (150/256)G + (29/256)B \\ C_b = (131/256)R - (110/256)G - (21/256)B + 128 \\ C_r = -(44/256)R - (87/256)G + (131/256)B + 128 \end{cases}$$

RGB to YCbCr formula of Equation (2)

$$\begin{cases} R = Y + 1.371(C_r - 128) \\ G = Y - 0.698(C_r - 128) - 0.336(C_b - 128) \\ B = Y + 1.732(C_b - 128) \end{cases}$$

III. QUANTUM PARTICLE SWARM ALGORITHM QPSO LOCATION

A. Principles and Process

QPSO also an evolutionary particle swarm algorithm, with "groups" and "evolution" concept, is also based on the individual (particle) size of fitness to operate [6,8-10]. QPSO each individual seen dimensional search space without a weight and volume particle and search space to a certain speed flight. The flight speed dynamically adjusted by the flying experience of individuals and groups of flying experience. Each particle represents a position-dimensional space, in two directions to adjust the position of a particle the following:

The optimal position of each particle has been discovered;

Best position particle swarm.

Each particle contains the following information:

$x_i = (x_{i1}, x_{i2}, ..., x_{id})$ The current position of the particles;

$v_i = (v_{i1}, v_{i2}, ..., v_{id})$ The current velocity of the particles;

$P_i = (P_{i1}, P_{i2}, ..., P_{id})$ It represents the best adaptation value of particles, namely pbest;

$P_g = (P_{g1}, P_{g2}, ..., P_{gd})$ Best adaptation value represents particle swarm that gbest.

Particle evolution equation is:

$$mbest = \frac{1}{M}\sum_{i=1}^{M} P_i = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1}, ..., \frac{1}{M} P_{id}\right) \quad (3)$$

$$p_{id} = \varphi * P_{id} + (1-\varphi) * P_{gd} \qquad \varphi = rand \quad (4)$$

$$x_{id} = p_{id} \pm \alpha * \left| mbest_d - x_{id} \right| * \ln(\frac{1}{u}) \quad (5)$$

Here is an intermediate position of the particle populations, and between a random point. QPSO expansion coefficient of contraction, which is an important parameter QPSO convergence, is generally preferable = (1.0-0.5) * (MAXITER-T) /MAXITER+0.5.Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. QPSO clustering model

QPSO each particle represents a cluster, the number of particles in the clusters specified by the user. Each particle structure is as follows: $x_i = (m_{i1}, ..., m_{ij}, ..., m_{iN_C})$, The m_{ij} represents the first i clustering center vector of the j particles in the cluster C_{ij}.

Following are some symbols:

The dimension of N_d is the dimension of every sample. N_c is the number of clustering centers (provided by the user), which is the number of clusters to be formed. z_p represents the sample, p represents the center vector of m_j, and j represents a collection of data vectors forming the clustering C_j.

Standard clustering Euclidean geometry is based on the distance between samples (Euclidean), that is calculated from the sample and the center of each cluster, assign the sample to the center of the smallest distance vector. The sample is calculated using the following formula $d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2}$ (6) where the subscript representatives from the cluster center dimension to. The goal of clustering is to cluster the distance between the bigger the better, the smaller the better from the inside of the cluster.

C. QPSO parameter setting procedure

Use PSO locate the license plate need to solve two major problems: the search space coding and fitness function value is determined. (1) The entire area is all pending license plate image search range, variable parameters including location (x, y) the sub-image area and the size (width, height). In order to reduce the size of the parameter space, fixed-size sub-image to be determined, (width, height) were taken as a possible license plate size (85 • 24), to be a quantum particle swarm optimization search is completed you can then use a smaller amount of calculation Local difference Method and precise projection method to extract the license plate.

(2) appropriate function to determine the fitness function calculated according to the following formula:

$$u_{ij} = 0,1 \quad u \in U_{c \times n} \quad \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n$$

$$\min J = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} f_j \left\| X_j - C_i \right\|^2 \qquad (7)$$

Specific process is:

1. The feature vector extraction.
2. Initialize (cluster centers, local optimization, global optimization).

For T = 1: MAXITER

3. clustering, equation (6) based on Euclidean geometry distance.
4. According to equation (7) calculated fitness function values
5. (3) calculated according to the formula.
6. Update the local optimum pbest.
7. Update the global optimum gbest.
8. (4) calculated according to the formula of random points
9. (6) Update Center vector particles according to the formula

end

Repeat steps 2 through 8 to calculate the number of iterations until satisfied. Texture feature vector to be determined according to the area to be acquired and texture feature vector plates can be obtained comparing the determined area is the possibility of the license plate area. An excellent style manual for science writers is [7].

IV. TEST RESULTS

Using the quantum particle swarm optimization algorithm, the image of the vehicle under a variety of lighting conditions to capture the actual toll station conducted experiment, through 100 real shot image positioning, positioning the image of 97 successful positioning success rate > 97%, positioning the failure of three images are caused by plate pollution, caused missing characters more. For some other location algorithm can not accurately locate the images contained in the test sample using a quantum particle swarm optimization positioning have very good results, as shown in the following figure. Which Figure 1 is a good case light shooting, clear images, the image of such obvious characteristics plate, various positioning algorithms are easier to locate, Figure 3 is a welcome headlights at night and when you capture an image, use such images Conventional localization algorithm difficult to successfully position. Quantum particle swarm optimization algorithm enhancement solves the problem for the sub-region of the image, positioning is very successful. Figure 2 is in the case of anti-color plates, license plates are generally blue and white or black, but there are some circumstances the license plate is black and white or white with yellow letters, because the use of one-dimensional filter group belonging to the edge feature detection, anti-color and still get to be well positioned.

Figure 1 Clear image

Figure 2 Anti-color image


Figure 3 Night image

V. CONCLUSION

Traditional algorithms for feature extraction, and when the target image is poor quality or have a greater degradation of the target image while a small percentage of the share in full, it is difficult for effective image enhancement, target characteristics for effective extraction. The quantum particle swarm optimization algorithm can be determined target area image full of small-scale image enhancement and extraction, avoiding the difficulty of dealing with interference or background full image quality caused by uneven experiment also proved that the method in the case of many poor image quality of the target area has a good target enhancement.

REFERENCES

[1] Wang Yu, Zhao zhenxiao, “Image region localization algorithm based on line edge recognition”, Computer Engineering 1999 (09):61-62.

[2] Zeng jianchao, Jie qian, Cui zhihua, “particle swarm algorithm” [M] Beijing: Science Press, 2004

[3] Liu Yong, Zhang ling, He wei, “Application of adaptive genetic algorithm in license plate location application”, Computer Applications 2008 (1):184-186.

[4] Ju Zhibin, “Application of Genetic Algorithm in License Plate giant Zhibin feature selection”, Computer Simulation 2010 (12):331-334.

[5] Liu qiaoge, Fu mengyin, Deng Zhihong, “Adaptive Algorithm ofLearning Rate for Feedfonlrard Neural Network”, System Simulation, 2006,18 (3): 698-705.

[6] Sun, J., Feng B., and Xu WB. “Particle Swarm Optimization with Particles Having Quantum Behavior”, Proceedings of 2004 Congress on Evolutionary Computation, 2004: 325-331.

[7] Fangfen Qi, “learning military information hiding technique based on binary images”, software, 2012,33 (1): 27-28

[8] Kennedy, J. and Eberhart, R. C. “Particle swarm optimization”, Proc. IEEE intel conf. on neural networks Vol.3, IEEE service center, Piscataway, NJ, 1995, 1942-1948.

[9] Eberhart, R. C. and Shi, Y. “Particle swarm optimization: developments, applications and resources”, Proc. congress on evolutionary computation 2001 IEEE service center, Piscataway, NJ., Seoul, Korea., 2001.

[10] J.Sun, and W.B.Xu, “A Global Search Strategy of Quantum-behaved Particle Swarm Optimization”, Proceedings of IEEE conference on Cybernetics and Intelligent Systems, 2004,111-116.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Foreground-background Segmentation Algorithm for Video Sequences

Zhou Wei
Dept. of Internet of Things Engineering
Jiangnan University
Wuxi, China
zhouwei.victoria@qq.com

Peng Li
Dept. of Internet of Things Engineering
Jiangnan University
Wuxi, China
pengli@jiangnan.edu.cn

HuangYue
school of IoT technology
Wuxi institute of technology
Wuxi, China
huangy@wxit.edu.cn

***Abstract*—Moving objects extraction is a crucial part of video surveillance system. This paper presents a foreground-background algorithm for motion detection. It is based on traditional adaptive mixture Gaussian model. By dynamically adjusting the parameters and the number of Gaussian components, the computation cost reduced greatly. In order to solve detected moving target based on Gaussian mixture model easily broken, two-way matching method on the basis of frame difference thoughts with a series of image filtering methods are combined. In a stable outdoor detector, the algorithm deals with lighting changes, swaying of leaves, and various noises reliably. The proposed algorithm can identify moving objects more exactly than traditional method. An intrusion detection alarming system can be built to discover the abnormity by using the algorithm to process and analyze the video sequences.**

Keywords—moving object detection; background extraction; goreground segmentation ; Gaussian mixture model; intrusion detection alarming

I. INTRODUCTION

The number of cameras available worldwide has increased dramatically over the last decade. In order to detect, segment, and track objects in videos automatically, several approaches are possible. Background subtraction is one of the most widely used methods in automatic video content analysis [1]. Numerous methods for background subtraction techniques have been proposed.

Wren et al. studied the statistical properties of the pixel value on time axis, and suggested Gaussian background model [2]. Stauffer and Grimson modeled the values of a particular pixel as a mixture of Gaussian. They proposed Gaussian mixture model (GMM) to deal with the multi-modal appearance of the background of a dynamic environment [3]. Zhang predicted the position of the moving object in the current frame using historical motion information, and then used the prediction information to construct a predictive model [4]. Yang Tao et al. presented a novel multiple layer background model to detect and classify foreground [5]. Cai Nian proposed a moving object detection method by combining a Gaussian mixture model with the wavelet transform to improve the robustness of the segmentation method [6]. Chen Shiwen proposed a weighted method based on the traditional learning rate in Gaussian mixture model updating step, which gives different weights to the mean value and the variance value correspondingly [7]. Zhang Heng et al. proposed an algorithm based on adaptive learning Gaussian mixture model by defining an efficiency factor between pixel samples and their background models [8]. The accumulation of efficiency factor shows how well the models can represent the background and is used to adjust the learning-rate dynamically.

Gaussian mixture model is probably the most popular background modeling technique which can describe the multi-model appearance of the background of a dynamic environment. However since its high calculation cost, the speed can't satisfy video surveillance system. Furthermore, the moving target detected by Gaussian mixture model is easily broken and with various noises [9]. To solve these two problems, this paper proposed a foreground-background algorithm based on improved Gaussian mixture model. By dynamically adjusting the parameters and the number of Gaussian components, the computation cost reduced greatly. Combining two-way matching method based on frame difference thoughts with a series of image filtering methods, the method can extract the moving objects exactly which is superior to the traditional method. Intrusion detection alarming system using this algorithm was built to discover the abnormity by processing and analyzing the video sequences, send the alarm and save useful information without human action.

II. BACKGROUND EXTRACTION

A. Gaussian mixture modeling

A mixture of adaptive Gaussian considers the value of a particular pixel over time as a "pixel process". At any time, t, the history of a particular pixel is $\{X_1,...,X_t\}$. The recent history of each pixel is modeled by a mixture of k Gaussian distributions:

$$P(X_t)=\sum_{i=1}^{k}\omega_{i,t}*\eta(X_t,\mu_{i,t},\Sigma_{i,t}) \quad (1)$$

where k is the number of Gaussian distributions; $\omega_{i,t}$ is the weight of the i^{th} Gaussian at time t, and $\sum_{i=1}^{k}\omega_{i,t}=1$; $\mu_{i,t}$ is the mean value. $\Sigma_{i,t}$ is the covariance matrix, which is assumed to be of the form $\Sigma_{i,t}=\sigma_{i,t}^2 I$. η is a Gaussian probability density function.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.92

1) Initialize:

The first Gaussian is initialized with a current pixel value as its mean value and has a high prior weight. While other mean values are zeros and weights are averaged. Variances of all Gaussians are equal.

2) Update:

Every new pixel value is checked against the existing K Gaussians, until it finds a match which is defined as

$$\left|X_t - \mu_{i,t-1}\right| < c * \sigma_{i,t-1} \tag{2}$$

The parameters of the Gaussian distribution which matches the new observation are updated as follows

$$\omega_{i,t} = (1-\alpha) * \omega_{i,t-1} + \alpha \tag{3}$$

$$\mu_{i,t} = (1-\beta)\mu_{i.t-1} + \beta X_t \tag{4}$$

$$\sigma_{i,t}^2 = (1-\beta)\sigma_{i,t-1}^2 + \beta\left(X_t - \mu_{i,t}\right)^T\left(X_t - \mu_{i,t}\right) \tag{5}$$

where α is learning rate; β is parameters updating rate:

$$\beta = \alpha\eta\left(X_t, \mu_{i,t}, \Sigma_{i,t}\right) \tag{6}$$

For unmatched distributions, parameters μ and σ remain the same, and the weights are changed as

$$\omega_{i,t} = (1-\alpha)\omega_{i,t-1} \tag{7}$$

If none of the k distributions matches the current pixel value, the least probable distribution is replaced by a distribution with the current value as its mean value, an initially high variance, and a low prior weight. At last, the weights of all Gaussian distributions are renormalized.

3) Order:

First, all Gaussian distributions are ordered by the value of $\omega_{i,t}/\sigma_{i,t}$. Then, the first K distributions are chosen as the background model:

$$K = \arg\min_k\left(\sum_{i=1}^{k}\omega_{i,t} > T\right) \tag{8}$$

where T is a measure of the minimum portion of the data that should be accounted for by the background.

B. Improved GMM

Traditional GMM deals with slow lighting changes, swaying branches, repetitive motions and other troublesome features of the real world, but at a cost of a prohibitive processing time that is unsuitable for real-time applications. There are three main approaches solving this problem.

1) Initial background

Traditional GMM uses the first frame as its initial background. However, in first frame of complex scene, there are some targets and noises which have influence on modeling. To solve this problem, statistical background based on probability was used as initial background.

First, n frames are selected. Accordingly, every pixel has n values. The pixel value with maximum probability of occurrence is used as initial value of background model. It can save some time of updating and get better results.

2) Parameters updating

In the phase of updating parameters, traditional GMM judgments and updates all pixels. Calculating covariance and updating rate cost a lot of time. It is far away from real time. To solve the problem, the algorithm is improved:

a) Simplify β : As (3) and (7) show, the weights of matched Gaussians become bigger and bigger as time goes by. When current model describes the background well, it can properly slow down the parameter updating rate. On the contrary, it is speed up. This paper changed the learning rate as $\beta = \alpha/\omega_{i,t}$, which not only updated Gaussian model better but also simplified the calculation and saved time.

b) Adaptive α : When there are large amount of foreground pixels, background model may be heavily affected and the learning speed should be accelerated. So learning rate α can be defined base on the number of foreground pixels in order to adapting the changing scene:

$$\alpha = \varepsilon(C_{t-2} + C_{t-1} + C_t)/3N \tag{9}$$

where C_t is the number of foreground pixels at time t, N is the total number of pixels, ε is a conversion constant.

3) *Adaptive GMM:*

For traditional GMM, the number of Gaussians is fixed. However in practical scene, state number of different regions is not the same. State number of a particular region may change as well. Keeping fixed number of Gaussians can waste a lot of time, so it is necessary to adaptively select the number of Gaussians according to the changing scene.

It can be seen from (3), (7) and (8), as time goes by, the weights of unmatched Gaussians will become smaller and smaller. These Gaussians will gradually become a part of Gaussians which represents foreground. Here the "expired" Gaussian is defined as

$$\omega_{i,t} < \omega_{init} \quad and \quad (\omega_{i,t}/\sigma_{i,t}) < (\omega_{init}/\sigma_{init}) \tag{10}$$

After sorting, the expired Gaussians will be behind the new initialized Gaussian. Obviously, the expired Gaussians not only cost a lot of time but also waste computing resources of the system, which should be deleted.

According to above analysis, adaptive Gaussians number selection strategy is put forward: 1) initialization: set only one Gaussian for each model in the scene; 2) updating: as the scene changes, when Gaussian mixture model doesn't match the current pixel value, if the number of Gaussians is smaller than the preset maximum value, add a new Gaussian, else the least probable distribution is replaced with the new Gaussian; 3) deleting: after updating the model every time, judge whether a Gaussian is expired or not, then delete the expired Gaussian.

III. FOREGROUND SEGMENTATION

After background extraction, preliminary background is formed and foreground targets can be detected. However, the detected targets are easily broken. There may be some misjudged targets as well. To detect the objects exactly, a series of subsequent has been carried out.

A. Background gradient difference

For outdoor environment, the adjacent pixels have similar temporal and spatial distribution. So it used neighborhood background subtraction:

$$BK(x,y) = \min_{i,j}[|\ f_t(x,y) - B_t(x+i,y+j)\ |] \qquad (11)$$

where $i, j \in \{-1,0,1\}$, f_t is the current frame, B_t is the real-time background. 3×3 neighborhood was selected here.

B. Neighborhood smooth

After the above process, it turned the difference image into a binary image. There were different kinds of noise in the binary image. 3×3 median filter was carried out to improve the quality of image and weaken the noise.

C. Morphology processing

After the above procedure, there may be some holes in foreground objects and some thin protrusions at the edge of foreground objects. Morphological image processing was applied to improve this situation. For removing the noise and saving details at the same time, this paper mixed the opening operation with closing operation which processing included twice erosion and twice dilation.

D. Filter in a single fame

Due to influence of light, wind and shadow, inevitably, there are some isolated noises affecting foreground. These noises can be removed by threshold method. This paper used Canny operator to detect the contours of connected components. First it counted the pixels in the contour. The number of these pixels was A (area of connected component). If $A < Threshold1$, the connected component was considered to be the noise that should be removed, else if $A > Threshold1$, it counted the pixels of the contour. The number of these pixels was P (perimeter of connected component). If $P > Threshold2$, the connected component was judged to be moving object, else if $P < Threshold2$, it removed the connected component. Thresholds can be adjusted by users according to monitoring environment.

E. Filter between frames

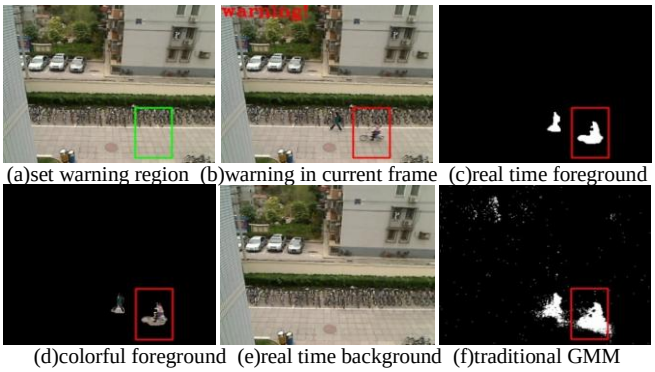
The small noise points have been generally removed. However, there may be some noise with bigger area. Since moving object is continuous in temporal domain yet noise is random in general, this paper proposed two-way matching method based on frame difference thoughts to filter the bigger discontinuous random noise [10]. Foreground mask of current frame and that of the next frame are operated by "and". If the area of matched region is larger than threshold Td, the foreground is saved, else if the area of matched region is smaller than threshold Td, it verifies current frame and the previous frame. If the area of forward matched region is larger than threshold Td, the moving object is considered disappearing, else discard the connected component. Threshold Td is defined as l times of maximum area of connected components in current frame, $l(0 < l < 1)$ can be adjusted according to the scene.

Through a series of processing, it can get background image and binary foreground mask image. The intact foreground image can be achieved by "and" operation of current frame and foreground mask image.

IV. Experimental Results

Intrusion detection alarming system using this algorithm was built to discover the abnormity. Main procedures are as follows: 1) acquire video images of monitoring scene; 2) set the warning region according to requirement and destination; 3) detect every frame of video sequences to get moving objects; 4) compare moving objects with warning region, if the object intrudes, send the alarm and save the current image; 5) turn all saved images into a short video. Program of the system was written in VS2010+ OpenCV2.2 environment, using PC hardware DELL (Pentium (R) Dual-Core CPU, 3GHz, 2GB memory).

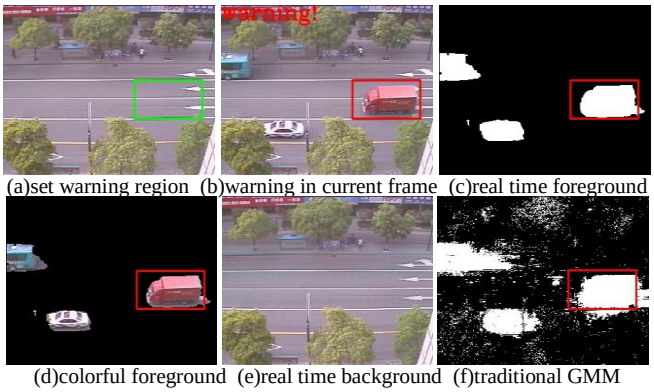
In order to test the performance of the algorithm, two videos were applied. As Fig.1 and Fig.2 show, foreground image detected by traditional GMM contains a lot of noise, and the moving targets are not accurate and complete. The proposed algorithm can reliably deal with various noises and exactly extract the moving objects with less false foreground pixels, which is superior to the traditional GMM.



(a)set warning region (b)warning in current frame (c)real time foreground

(d)colorful foreground (e)real time background (f)traditional GMM

Figure1. Campus scene detection results



(a)set warning region (b)warning in current frame (c)real time foreground

(d)colorful foreground (e)real time background (f)traditional GMM

Figure2. Road scene detection results

For quantitative analysis of the experimental results, the proposed algorithm was compared with traditional GMM, [11] and [12]. Different algorithms are evaluated and compared in conditions of the same hardware and software. The most widely used metrics to assess the performance of background-foreground segmentation are Recall and Precision [13]. F metric is the weighted harmonic mean of Precision and Recall [14], it is defined as

$$F = 2 \times R \times P / (R + P)$$

As Tab.1 and Tab.2 show, both precision and recall of this algorithm are higher than those of GMM, [11] and [12]. Due to filling holes in connected components, the detected targets are accurate and the recall is higher than other algorithms. Through a series of subsequent processing, the algorithm can overcome the affect of leaves swaying and slight light changes, filter out most of the noise in the image, and eliminate the false foreground information. The precision has been greatly improved.

In order to test the time efficiency of the algorithm, this algorithm was compared with GMM, [11] and [12]. Tab.3 indicates the traditional mixture Gaussian background modeling method is time-consuming. The algorithm in [11] simplified the calculation, so the speed is very fast. However, it is at the expense of the decrease of recall. To obtain high detection rate, the algorithm in [12] needs to construct foreground model and calculate the short-time stability, which results in high computational complexity and long time consuming. Due to the improvement of updating parameters and adaptive selection of the number of Gaussian components, the algorithm this paper proposed reduces the amount of calculation greatly to meet the fast processing that monitoring system needs.

From above, the proposed algorithm can detect invasive targets well, send the alarming, and save the current frame when there are moving objects intrude the warning area. Compared with the original surveillance video, the length of new video is greatly reduced which is convenient for regulators to check. By modeling the warning area only, it can achieve real-time monitoring. Applying this method to video monitoring system has great practical value.

TABLE1. Experimental result of different algorithms in campus scene

Background model	Recall	Precision	F metric
GMM	0.77	0.62	0.69
Reference[11]	0.81	0.73	0.77
Reference[12]	0.85	0.74	0.80
This paper	0.89	0.82	0.85

TABLE2. Experimental result of different algorithms in road scene

Background model	Recall	Precision	F metric
GMM	0.72	0.58	0.64
Reference[11]	0.78	0.69	0.73
Reference[12]	0.80	0.73	0.76
This paper	0.83	0.77	0.80

TABLE3. Test result of single frame processing time (t / ms)

Algorithm	Campus	Road traffic	Average
GMM	84	78	81
Reference[11]	67	61	64
Reference[12]	75	68	72
This paper	61	55	58

V. CONCLUSION

The proposed foreground-background segmentation algorithm for video sequences dealt with slow lighting changes by slowly adapting the values of the Gaussians. It also dealt with multi-modal distributions caused by shadows, swaying branches, and other troublesome features of the real world. By dynamically adjusting the parameters and the number of Gaussian components, the computation cost reduced greatly. Combining two-way matching method based on frame difference thoughts with a series of image filtering methods, the method can extract the moving objects exactly which is superior to the traditional method. The foreground-background segmentation algorithm this paper proposed was realized in Microsoft Visual Studio. It can extract the foreground information of different types quickly and accurately, and remove various noises effectively at the same time. The realized intrusion alarming system has excellent performance and obvious practical value.

REFERENCES

[1] Arseneau, Cooperstock J. Real-time image segmentation for action recognition[C]. Communications, Computers and Signal Processing, 1999: 86-89.

[2] Wrren C, Azarhayejani A, Darrell T. Pfinder: real-time tracking of the human body[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997: 780-785.

[3] Stauffer C, Grimson W. Adaptive background mixture models for real-time tracking[C]. IEEE CVPR, 1999: 246-252.

[4] Zhang Xiang, Yang Jie. A novel algorithm to segment foreground from a similarly colored background[J]. International Journal of Electronics and Communications, 2009, 63(11): 831-840.

[5] Yang Tao, Li Jiang, Pan Quan. A multiple layer background model for foreground detection[J]. Jounal of Image and Graphics, 2008, 13(7): 1303-1308.

[6] Cai Nian, Chen Shiwen, Guo Wenting. Moving object detection using Gaussian mixture model and wavelet transform[J]. Jounal of Image and Graphics, 20011, 16(9): 1716-1721.

[7] Chen Shiwen, Cai Nian, Tang Xiaoyan. Improved moving object detection algorithm based on Gaussian mixture model[J]. Modern Electronic Technology, 2010, 2:125-128.

[8] Zhang Heng, Hu Wenlong, Ding Chibiao. Adaptive learning Gaussian mixture models for video target detection[J]. Journal of Image and Graphics, 2010, 15(4):631-636.

[9] Chen Mingsheng, Liang Guangming, Sun Jixiang. Fast moving object detection method using temporal-spatial background model[J]. Journal of Image and Graphics, 2011, 16(6):1002-1007.

[10] Neri A, Clonnese S, Gusso G, et al. Automatic moving object and background separation[J]. Signal Processing, 1998: 219-232.

[11] Jiang Peng, Qin Xiaolin. Robust foreground detection with adaptive threshold estimation[J]. Journal of Image and Graphics, 2011, 16(1): 37-43.

[12] Zhang Chao, Wu Xiaopei, Zhou Jianying. A moving object detection algorithm based on improved GMM and short-term stability measure[J]. Journal of Electronic and Information, 2012, 34(10): 2402-2408.

[13] Maddalena L, Petrosino A. A self-organizing approach to background subtraction for visual surveillance applications[J]. IEEE Transactions on Image Processing, 2008, 17(7):1168-1177.

[14] Gan Chao, Wang Ying, Wang Xiangyang. Multi-feature robust principal component analysis for video moving object segmentation[J]. Journal of Image and Graphics, 2013, 18(9): 1124-1132

Processing of Words Labels in Scanned Map Based on Singularity Detection

Xu Zhipeng
School of Physics Science and Information Engineering,
Liaocheng University
Shandong Provincial Key Laboratory of Optical Communication
Liaocheng, China
xuzhipeng@lcu.edu.cn

Liu Runqing
Shandong Linqing polytechnical Vocational School
Linqing, China
Gyxxjdjyz@126.com

***Abstract*—The words labels in the scanned map should be removed in some application. In this paper singularity detection based on wavelet is adopted to separate the words labels from the scanned map. One row is extracted from the scanned map, and then is convoluted with wavelet in different scales. Compared to the end of words labels, the words labels themselves have different Lipschitz exponents. In this paper the local minimum of Lipschitz exponents is proposed to separate the words labels from the background.**

Keywords-image processing; singularity; wavelet

I. INTRODUCTION

The geographic information system(GIS) extract information from scanned map, and recognized as different data object. Many maps are recorded on papers. When these maps are scanned and input into the computers, the geographic information, such as edges of region, streets, etc, should be accurately separated. Words labels exist in most maps. In some application, such as computation of area or circumference of specific region, these words labels should be eliminated. Traditional methods are edge detection and morphological operations. There are some problems in these methods, such as double edges, discontinuity and too many edges to analyze. In this paper with careful observation of one row data in scanned map, words labels have different characters to the other regions. The difference can be characterized by singularity with wavelet transform.

II. APPLICATION OF WAVELET TRANSFORM

Wavelet transform has a wide application in image processing, power supply system, diagnosis of mechanical equipment, etc. Mallat and Zhong[1] study the property of mutliscale edges of an image with wavelet theory, the edges can be described by local maxima of wavelet transform. Mallat and Hwang[2] explain the singularity with Lipschitz exponent, proved that the irregular signal can be detected by local maxima.

The detection of signal singularity with wavelet transform has widely application in power supply system. Usually the voltage and current in power supply system are regular periodic function. When an abnormal change happens, the amplitude of voltage may get a transient drop. Both the amplitude and the time should be recorded accurately. In reference [3] the coefficients of wavelet transform are used to locate the time and amplitude of voltage drop. The fault diagnosis of gearbox can be realized by singularity. ZHU Zhong—kui [4] proposed that the signal transient of gearbox can be extracted by selecting proper threshold of reconstruction of transient feature.

In a scanned map, the words labels behave like an abnormal change. The other regions of image also have gray level changes, such as the edge of different regions. In this paper the different part of words labels are analyzed with wavelet transform. Different parts of the words labels have different singularity. The local minimum is used to detect the words labels.

III. WAVELET TRANSFORM

The wavelet transform was introduced by Morlet 和 Groddmann[5]. Let $\psi(x)$ be a complex valued function, $\psi(x)$ is called a wavelet when it satisfies the following equation:

$$\int_{-\infty}^{+\infty}\psi(x)dx = 0$$

Let $\psi_s(x)=(1/s)\psi(x/s)$ be a dilation of $\psi(x)$ at scale s, the wavelet transform of function $f(x)$ is defined as:

$$Wf(s,x)= f * \psi_s(x)$$

The symbol * means convolution of two functions. The Lipschitz exponent is defined as[2] :

Let n be a positive integer, $n \le \alpha \le n+1$, a function $f(x)$ is said to be Lipschitz α ,at x_0 ,if and only if there exists two constants A and $h_0 > 0$, and a polynomial order n, $P_n(x)$,such that for $h < h_0$:

$$|f(x_0 + h)- P_n(h)| \le A|h|^{\alpha}$$

For a function like Dirac at $x = 0$, the Lipschitz exponent is -1. While the Lipschitz exponent of a step function is zero. In reference [6] a theorem is proposed to describe the connection of Lipschitz exponent and wavelet transform:

Let $f(x)\in L^2(R)$, $[a,b]$ is an interval of real numbers, $0<\alpha<1$, for any $\varepsilon>0$, $f(x)$ is uniformly Lipschitz α over $(a+\varepsilon,b-\varepsilon)$, if and only if there exists a constant A_ε, such that for $x\in(a+\varepsilon,b-\varepsilon)$ and s>0:

$$|Wf(s,x)| \le A_\varepsilon s^{\alpha}$$

Also, there is a theorem of Lipschitz exponent at a point [7]:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.93

Let n is an integer, $\alpha \le n$, $f(x) \in L^2(R)$, if $f(x)$ is Lipschitz α at x_0, then there exists a constant A such that for all points x in the neighborhood of x_0 and any scale s,

$$|Wf(s,x)| \le A\left(s^{\alpha} + |x - x_0|^{\alpha}\right)$$

Conversely, let $\alpha < n$ be a non integer value. The function $f(x)$ is Lipschitz α at x_0, if the following two conditions hold:

- There exists some $\varepsilon > 0$ and a constant A, such that for all points x in neighborhood of x_0 and at any scale
$$|Wf(s,x)| \le As^{\varepsilon}$$
- There exists a constant B, such that for all points x in neighborhood of x_0 and at any scale
$$|Wf(s,x)| \le B\left(s^{\alpha} + \frac{|x - x_0|^{\alpha}}{|\log|x - x_0||}\right)$$

In order to explain the above two conditions, the cone of influence is introduced. The influence cone usually is used to explain the relationship between the wavelet and Lipschitz exponent. In figure 1 under scale space of $[s,x]$, the influence cone of a point x_0 is all the points that satisfy[2]:

$$|x - x_0| \le Cs$$

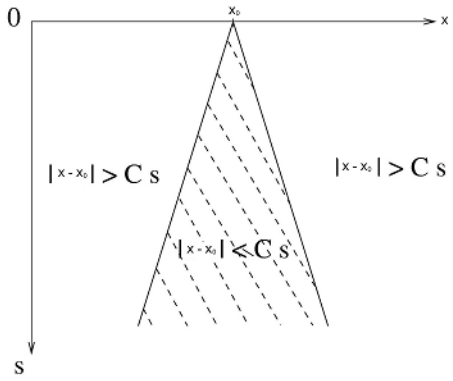


Figure 1. The cone of influence

Mallat[8] points out that if the derivative of nth order of function $f(x)$ has no oscillation in the neighborhood of x_0, then the Lipschitz exponent of point x_0 is controlled by the wavelet transform in the cone of influence.

Lipschitz exponents can be measured by the slope of decay of wavelet transform maxima at fine scales in log coordinate [2]. Suppose there exists a scale $s_0 > 0$ and a constant C, such that for $x \in (a,b)$ and $s < s_0$, all the modulus maxima belongs to a cone defined by $|x - x_0| \le Cs$, then the function $f(x)$ is said to be Lipschitz α ,at x_0 , if and only if there exits a constant A, such that all the modulus maxima in the cone satisfy:

$$|Wf(s,x)| \le As^{\alpha+1/2}$$

The following results are simulated on the wavelab850 issued by Stanford University [9].

IV. Process of Scanned Map

Figure 2 is an administrative map of Shandong province; there are many words labels on the map. These words clearly label the different regions of Shandong province. At the same time the words labels also overlap some edges of different regions. In some images processing applications, such as the calculation of areas of different cities, these labels should be removed.

Figure 2. Administrative map of Shandong Province

The color image has plenty of information. These color information help us segment different objects; at the same time the color information makes the computation expensive. So the color image is usually converted into gray tone image with the following equation:

Gray value = 0.3*R + 0.59*G + 0.11* B

The RGB represent the red, green, and blue component of the color image. Figure 3 is a converted gray image. The black line at the ordinates 370 is the line that will be discussed. There are some words labels in line 370, such as "jinan" and "qingdao". Also there are some edges of different regions in the line. The figure 4 displays the No. 370 row data of figure 3. One can see the amplitude of row 370 varies from 0 to 256. This will cause plenty of computation. Usually the 256 gray levels is quantized to 16 levels, figure 5 is a quantized data of row 370. The irregularity remains, while the computation becomes simple.

Figure 3. Gray image of administrative map of Shandong Province

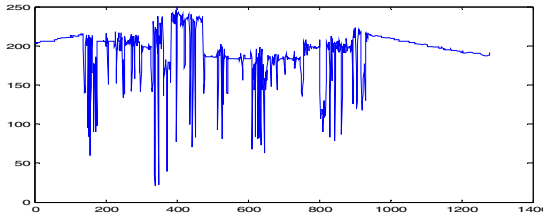


Figure 4. The No.370 row data from figure 3

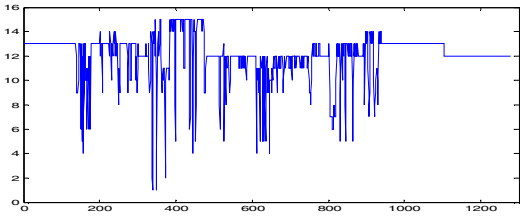


Figure 5. Quantized data of figure 4

In order to carefully discuss the signal, part of row 370, corresponding to the image is displayed in figure 6.

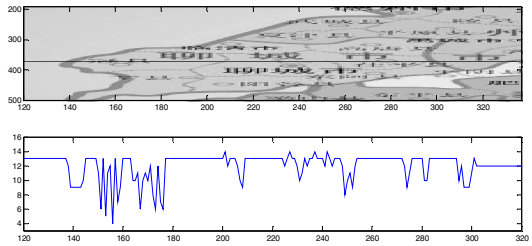


Figure 6. part of row 370 and corresponding image(start abscissa is 120)

From the abscissa 149 to 176, there is a word label "Guanxian", the signal varies from 4 to 13. While from the abscissa 177 to 199, that is the end of the word label, the signal does not change, the corresponding part of image is the blank region. The characters of these two parts of signal are different.

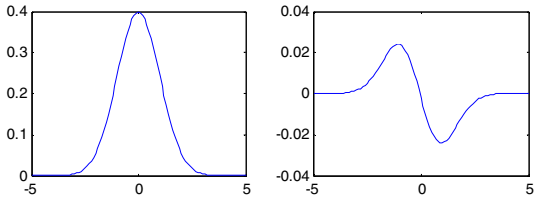


Figure 7. Gaussian function and its first derivative

The first derivative of a Gaussian is selected to be the wavelet function. Figure 7 plots the Gaussian function and first derivative. The signal in figure 6 is convoluted with the wavelet function. The modulus maxima lines of the wavelet transform is shown in figure 8. The neighborhood pixels of the words "Guanxian" cover from the abscissa 130 to 194(the displayed abscissa have been subtracted 130). The No.7 and No.8 maxima lines correspond to the word label "Guanxian", while the No.2 maxima line that converges to the point of abscissa 46 corresponds to the end of word label "Guanxian".

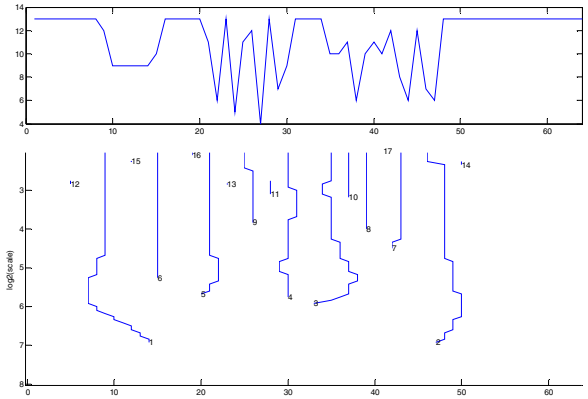


Figure 8. modulus maxima of wavelet transform of part of row 370

Figure 9 is the slope curve in log scale corresponding to No.2,No.7,No.8 modulus maxima lines in figure 8. Also two straight line are drawn with slope = 0.5, and slope = 0. One can see the No.2 maxima line coincides with the straight line with slope = 0, while the No.7,No.8 maxima lines coincides with the straight line with slope = 0.5. This means the point of abscissa 46 which No.2 maxima line converges to is Lipschitz -0.5, while the point of abscissa 39 which No.8 maxima line converges to is Lipschitz 0. The point of abscissa 46 has more singularity than the point of abscissa 39. If an appropriate threshold, for instance 0.2, is chosen to separate these two kinds of points, then the words labels can be separated from the image.

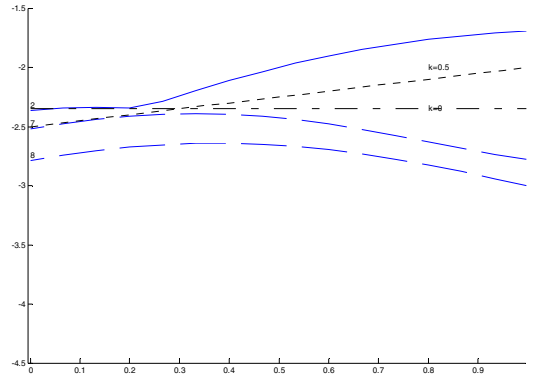


Figure 9. the slope in log scale corresponding to No.2,No.6,No.7 modulus maxima lines in figure 8

Figure 10 is another instance, the last word of word label "Jinan Shi", that is the word "Shi". The abscissa is from 436 to 455. The maxima lines are shown in figure 11. The No.7 maxima lines correspond to the word label "Shi", while the No.1 maxima line that converges to the point of abscissa 34 corresponds to the end of word label "Shi". Figure 12 is the slope in log scale corresponding to No.1, No.7, No.9 modulus maxima lines in figure 11. Also two straight lines are drawn with slope = 1.2, and slope = 0.9. One can see the No.1 maxima line coincides with the straight line with slope = 0.9, while the No.7, No.9 maxima lines coincide with the straight line with slope = 1.2. This means the point of abscissa 34 which No.1 maxima line converges to is

Lipschitz 0.4, while the point of abscissa 27 which No.9 maxima line converges to is Lipschitz 0.7. The point of abscissa 34 has more singularity than the point of abscissa 27. If an appropriate threshold, for instance 0.5, is chosen to separate these two kinds of points, then the word labels can be separated from the image.

From the above analyses of two words labels we realize that the threshold value of Lipschitz exponent is alterable; it is the local minimum of slope. In fact, the end part of word label behaves like a step function approximately; it is discontinuous at that point. Lipschitz exponent at this point is zero. While the word label itself behaves like a continuous curve approximately. The step function has more singularity than the continuous curve. The result accords with the singularity theory.

V. CONCLUSION

In this paper, the singularity of words labels in scanned map is discussed. One row data is extracted from the map. The row data is then convoluted with a wavelet function. With singularity theory, at the end part of words labels the singularity is more than the labels itself. The local minimum of Lipschitz exponents of the words label and the end part of words label can be used to separate the words labels.

REFERENCES

[1] S. Mallat, S. Zhong. Characterization of signals from multiscale edges. IEEE Transactions on Pattern Analysis and Machine Intelligence,1992, 14(7):710-732.

[2] S. Mallat ,W. L. Hwang.Singularity detection and processing with wavelets.IEEE Trans. Inform. Theory, 1992,38(2):617-643.

[3] Tunaboylu N S, Collins E R. The wavelet transform approach to detect and quantify voltage sags.Proc. of the 1996 IEEE Int. Conf . On Harmonics and Quality of Power,1996,619～624.

[4] ZHU Zhong－kui , KONG Fan－rang , WANG Jian－ping, etc. Detection and Extraction of Signal Transients and Their Applications in Fault Diagnosis, Journal of University of Science and Technology of China, 2004,Vol.34,1:105-110.

[5] A.Grossmann ,J.Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape, SIAM J. Math.,1984,15:723-736.

[6] M. Holschneider ,P. Tchamitchian.Regularite locale de la fonction non-differentiable de Riemann.Lecture notes in Mathematics, P.G. Lemarie,Ed. New York: Spring-Verlag, 1989.

[7] S. Jaffard,"Exposants de Holder en des points donnes et coefficients d'ondelettes," Notes au Compte-Rendu de l'Academie Des Sciences,France,1989,vol.308,ser.I,pp.79-81.

[8] Stephane Mallat. A wavelet tour of signal processing, 2nd Edition. Academic Press,1999.

[9] Jonathan Buckheit, Shaobing Chen, etc.Wavelab850, Stanford University,2005.

http://statweb.stanford.edu/~wavelab/

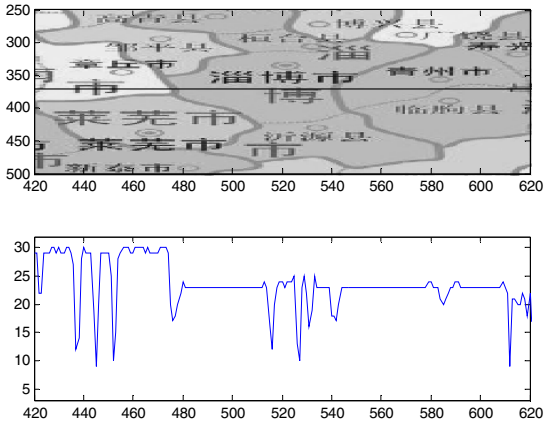


Figure 10. part of row 370 and corresponding image(start abscissa is 420)

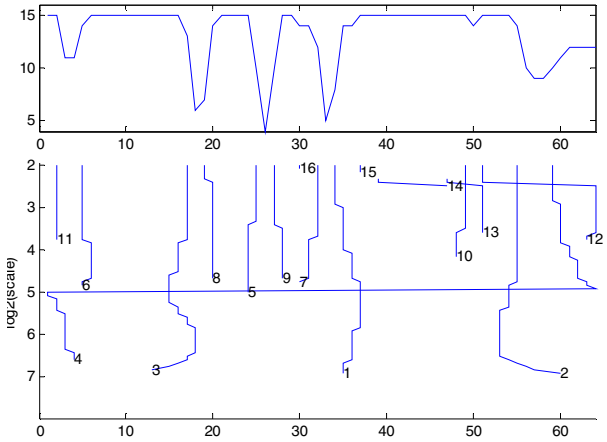


Figure 11. modulus maxima of wavelet transform of curve in fig.10

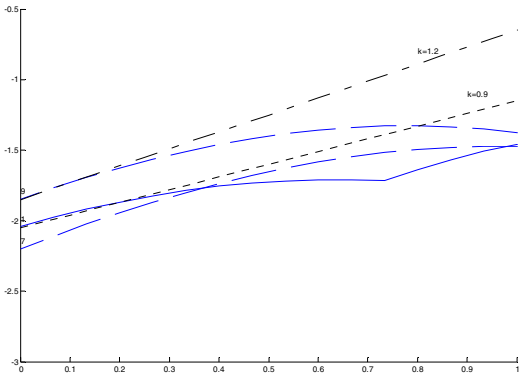


Figure 12. the slope in log scale corresponding to No.1,No.7,No.9 modulus maxima lines in figure 11

An Improved Live-wire Freed from the Restriction of the Direct Line between Seed Points

Zhou Di
Soochow University
Suzhou, China
e-mail: zhoudiprivate@msn.com

Xu Wenbo
School of Internet of Things Engineering,
Jiangnan University
Wuxi, China
e-mail: xwb_sytu@hotmail.com

***Abstract*—In the original live-wire algorithm, optimality is defined as the minimum cumulative cost path from a seed point to another, where the cumulative cost of a path is the sum of the local edge costs on the path. Therefore, the optimal path between two seed points is restricted not far away from the straight line between these two points. Consequently, when delineating a boundary with acute concaves and convexes, more seed points are required which can become very time consuming to calculate. This paper proposed an improved live-wire algorithm where the optimality is redefined as the minimum average cost path between two seed points, which frees the optimal path from the restriction of the straight line. In theory, the original live-wire algorithm is only a special case of our improved algorithm. Furthermore, the Canny edges on the boundary of the target object is weighted according to the gray or color value of the seed point. Experiments conducted on a variety of image types have shown that this improved live-wire algorithm requires less seed points than the original algorithm when delineating the same boundary and as a result, reduces the time required to complete the calculation..**

Keywords- segmentation; live wire; average cost path; weighted Canny edge

I. INTRODUCTION

One main challenge of performing automatic image analysis is the task of separating various individual objects. In practical applications, images are usually distorted by noises and inter-camera variability, making segmentation of the images more difficult. Although fully automatic segmentation schemes can significantly reduce the time demand, human validation is still required to guarantee its accuracy. Therefore, semi-automatic methods such as live-wire[1, 2], snake[3, 4] and varieties of active contour model[5-8] were proposed to obtain high accuracy, efficiency and reproducibility rates. The live-wire starts with a manually specified seed point on the boundary of the object edge, followed by calculating the minimum cost of reaching every point on the image from the given seed point. The path that produces the minimum cost between two points in the boundary is likely to be along the edge of the object of interest because the cost terms are designed so as to give edges low costs. Dijkstra's algorithm is generally used to find this path with the minimum cost[9]. Therefore, when the mouse position moves close to an object edge, a "live-wire" boundary snap on and wraps around the object of interest. Input of a new seed point "freezes" the previous selected boundary segment, and the process is repeated until the boundary is complete. However, since calculation of the minimum cost to each point on the image from the seed point is the most time-consuming part of the live-wire algorithm, the improvement of the delineation accuracy by providing a more dense set of seed points is at the expense of segmentation speed.

In this paper, we propose a average cost calculation approach in order to overcome the shortcomings of the original optimal path searching method. The method severs the relationship between the costs of one path and its length, which in turn effectively reduces the number of seed points required in the segmentation process. In addition, new local component cost function is used to weight target edges, so as to prevents the "live-wire" boundary from snapping onto the non-target edges when moving the mouse..

II. METHODS

Boundary searching via live-wire could be described as the process of finding the globally optimal path from a start node to a goal node, which consists of two important issues to address: how to define the criteria for "optimal path", and how to find this optimal path.

A. Cost terms for gray images

Generally, for an image I, the cost function is a weighted sum of the component cost items for those features corresponding to the object boundaries. Features widely used by the existing works include the gradient magnitude f_G, gradient direction f_D, Canny edge detection f_C, and Laplacian zero crossing edge detection f_Z[1, 2, 10]. The cost on the directed link from node p to a neighboring node q is calculated by

$$C(p,q) = w_Z \cdot f_Z(q) + w_C \cdot f_C(q) + w_G \cdot f_G(q) + w_D \cdot f_D(p,q) \quad (1)$$

where w_Z, w_C, w_G, and w_D are the weight constants used to control the contribution rate of corresponding cost terms in the total cost. For gray images, the cost term for the gradient magnitude of node q is defined as

$$f_G(q) = 1 - G(q) / \max(G) \quad (2)$$

where $G(q)$ is the magnitude of the image gradient and $\max(G)$ is the highest gradient of the entire image. In order to make the high image gradients correspond to low costs, this term should be subtracted from 1. The cost term for the gradient direction from node p to node q is defined as

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.94

$$f_D(p,q) = \arccos(D_x(p)/G(p) * D_x(q)/G(q) + D_y(p)/G(p) * D_y(q)/G(q))/\pi \quad (3)$$

where $D_x(p)$ and $D_y(p)$ denote the horizontal and vertical gradients of node p, derived by convolving the horizontal and vertical Sobel operators, respectively.

Both Canny and Laplacian zero crossing edge detections are important methods of generating binary edges for images. The outputs of edge detection by either of these two methods are in the form of binary images with edges represented by 1 and the rest of the background pixels containing 0. Similarly, these binary images are also need to invert for producing strong edges with low costs. However, Canny method produces many spurious edges while detecting the edges that are on the boundary of the target object, and these spurious edges could often cause misguided cost calculation. Therefore, in order to prevent the live-wire snapping to undesired edges, we weight those Canny edges on the target boundaries, but weaken the others relatively.

n_{11}	n_{12}	n_{13}	••	n_{1L}
n_{21}				
n_{31}		s		
⋮				
n_{L1}	n_{L2}	n_{L3}		n_{LL}

Figure 1. The square region used to calculate the average gray value of neighbors of s.

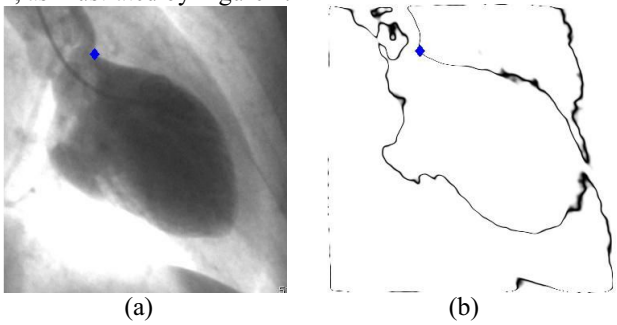
Notice that, each segment of live-wire boundary is generated from a seed point which is usually on or near the boundary of the target object. Therefore, the average gray value of the square region with seed point s as its center and side length L can be easily calculated as:

$$av\_gray(s) = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} I(n_{ij}) \quad (4)$$

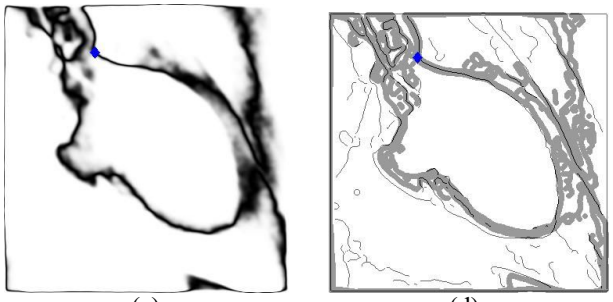
where n_{ij} represents those neighbor nodes, as shown in Figure 1. Then a simple Gaussian Filter could be used to emphasize the regions with the similar gray value of $av\_gray(s)$ in image I:

$$I' = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(I-av\_gray(s))^2}{2\sigma^2}} \quad (5)$$

where different σ results in different emphasized images I', as illustrated by Figure 2.



(a) (b)



(c) (d)

Figure 2. Emphasize the specific pixels with the similar gray value as the seed point's, (a), the original DSA image with the seed point in blue diamond, (b), the emphasized image by σ=2, (c), the emphasized image by σ=10, (d), superposition of the Canny edge of (a) and the dilated Canny edge of (c).

It is obvious that the Canny's edge detection on I' (*Canny_I'* from here), is much less than those detected on I. What's more, most of *Canny_I'* are near the real object boundary we want. However, as pointed above, I' is the image in which specific region with specific gray value emphasized, and therefore, the *Canny_I'* is almost but not all on the real boundary. Since our goal is to emphasize those Canny edges representing the real object boundary detected on I (*Canny_I* from here), the *Canny_I'* could be dilated to cover more than single pixel width, which can be referred as the emphasized possible-boundary region f_{CE}. As shown by Figure 2(d), on the real heart boundary, Canny I and the emphasized possible-boundary region f_{CE} overlap with each other.

In conclusion, the cost of the directed link from node p to a neighboring node q in this paper could be calculated by

$$C(p,q) = w_Z \cdot f_Z(q) + w_C \cdot f_C(q) + w_G \cdot f_G(q) + w_D \cdot f_D(p,q) + w_{CE} \cdot f_{CE}(s) \quad (6)$$

B. Cost terms for color images

For color images, Equation (6) is also used as the cost function in this work. Different with gray images, the gradient magnitude of node q is usually defined as the maximum value among the gradient magnitudes for R, G and B color values [10], that is:

$$G(q) = \max\{G_R(q), G_G(q), G_B(q)\} \quad (7)$$

Moreover, all the neighbor nodes of seed point s should be divided into two categories according to the R, G, B color values, which represent outside and inside of the target boundary. The average color values of these two categories can be written as *av_color1(s)* and *av_color2(s)*. Therefore, the similarity between node q and the seed point s could be calculated by:

$$dis(q,s) = \sqrt{(av\_color1(s) - av\_color1(q))^2 + (av\_color2(s) - av\_color2(q))^2} \quad (8)$$

The smaller the *dis(q,s)* is, the more similar the node q is to seed point s, and with more possibility that the node q is on the real target boundary. The same as what done to the gray images, we can use a Gaussian Filter to emphasize those regions have a high possibility representing the real boundary:

$$I' = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-dis(I,s)^2}{2\sigma^2}} \quad (9)$$

The subsequent processing is the same as that for gray images: detecting Canny edge on *I'*, then dilating to form the emphasized possible-boundary f_{CE}. Figure 3(a) is a flower image with the seed point s in blue diamond. It could be noted that this flower has rich colors which make it has complex Canny edges. Figure 3(b) demonstrates the superposition of the Canny edges and the emphasized possible-boundary region $f_{CE}(s)$, which could be seen that those Canny edges on the real flower boundary are emphasized by $f_{CE}(s)$.

(a) (b)

Figure 3. Emphasized possible-boundary $f_{CE}(s)$ in color images, (a), the original flower image with seed point in blue diamond, (b), the superposition of the $f_{CE}(s)$ and the Canny edge of (a).

C. *Proposed Improvement of Live-wire Operations*

In original Live-wire, the costs of points on one path are calculated through accumulation in reverse order. For two points on the same path, the point closer to the beginning always bears lower cost than the point further away. In other words, the cost of one point is proportional to the path length.

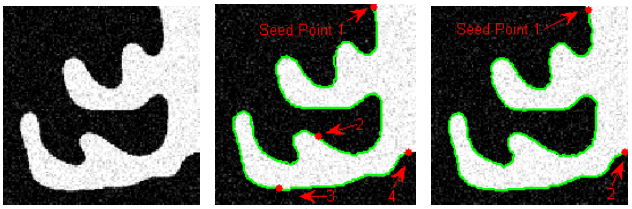
One solution is to eliminate the influence of path length on the point cost. Therefore, we can redefine the cost of point as the average cos. In this way, for the points on the same path, the cost of those far away from the start seed point is no longer always higher than those points near the beginning of the path. Therefore, the original Live-wire 2D graph searching algorithm can be improved as follows:

```
        s                    {Start seed point}
        l(q,r)    {Local cost function for link between point q and r}
    Data Structures:
        L                    {List of active points sorted by total cost. Initially empty}
        N(q)                 {Neighborhood set of q (8 neighbors)}
        e(q)                 {Boolean function indicating if q has been processed}
        g(q)                 {Cost from start seed point s to q}
    Output:
        P(q)                 {Prior point of q in the minimum cost path}
    P_N(q)                   {Number of points on the path from start seed point s to q}
    g_av(q)                  {Average cost from start seed point s to q}
    Improved Live-wire 2D Graph Searching Algorithm:
        g(s)←0; L←s; g_av(s)    ←0;
                             {Initialization }
        while L is not empty do begin
        {While still points to expand}
                q←min(L);
                {Remove minimum cost point q from active list}
    e(q)←TRUE;                                   {Mark q as expanded}
    for each r∈N(q) such that not e(r) do begin
        gtmp←g(q)+l(q,r);                                ( * ) {Compute total cost of neighbor r}
        g_avtmp←gtmp / P_N(q)                            ( ** ) {Compute average cost of r}
        if  r∈L and g_avtmp <g(r) then
                r←L;
        {Remove higher cost neighbor’ s from list }
        if  r L then begin
                g(r)←gtmp;
        {Assign neighbor’ s total cost,}
                g_av(r)←g_avtmp;
        {Assign neighbor’ s average cost,}
                P_N(r)←P_N(q)+1;
        {Count the path length from s to r}
                P(r)←q;
        {Set r’ s prior point as q}
                L←r;
        {Put r into active list}
        end
    end
end
```

III. RESULTS

In order to demonstrate the advantages of the proposed algorithm, we tested it on a variety of image types compared with the original live-wire. Figure 4 is an artificial test image that exhibits edge blurring and Gaussian noise. Figure 5 is a CT image of a spinal vertebra. Since flower recognition is widely researched recent years [11-13], we also test our proposed algorithm on a colorful flower image as Figure 6. The edge detection starts from *Seed Point 1* on every image in a clockwise fashion. All the subsequent seed points are inputted when necessary, which means the “live wire” boundary will no longer snap to the target object as the mouse moves on.

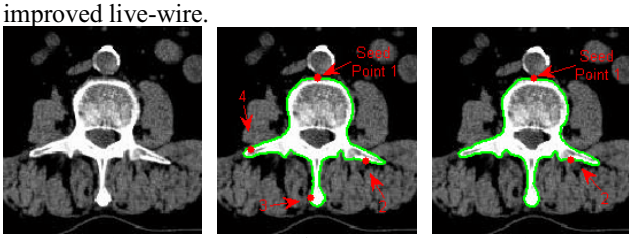
Figure 4 belongs to those images with simple background, which can highlight the superiority of using average cost instead of cumulative cost, by limiting the difference in cost terms of between the original and improved Live-wire. Figure 4 contains several saw-teeth with edge blurring and Gaussian noise. Searching the same boundary, the original Live-wire needs 4 seed points in total, including the start point and the end point, while the improved Live-wire with k=1 is able to snap the boundary tightly from the start to the end, without any additional seed point in the midway.



(a) (b) (c)

Figure 4. (a), Artificial test image with edge blurring and Gaussian noise, (b), Boundary detection using the original Live-wire, 4 seed points totally needed, (c), Boundary detection by the improved Live-wire, k=1. No need any additional seed point between the start point and the end point.

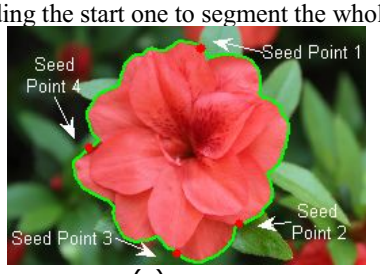
Figure 5 is a CT image of a spinal vertebra. The end point is the same as the start point: both are the *Seed Point 1*. It could be noted that three additional points are needed when the original live-wire is used, while only one seed point is inserted during the boundary detection by the improved live-wire.



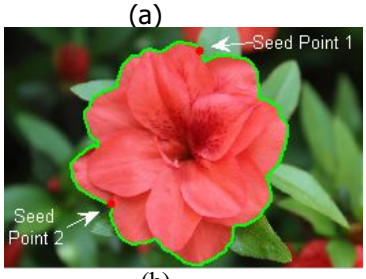
(a) (b) (c)

Figure 5. (a), Spinal vertebrae, (b), boundary detection utilizing the original Live-wire, 3 additional points needed, (c), using the improved Live-wire (k=1), only 1 seed point should be inserted to define the entire boundary of the spinal vertebrae.

Figure 6 is an image visualizing a colorful flower with multi-layered petals and green-leaf background. Searching the boundary of the image with Canny method, we may extract many non-real boundaries. With *seed point 1* the start point, the original Live-wire algorithm needs other three points to outline clockwise the whole border of the flower, while the improved Live-wire algorithm only needs two points including the start one to segment the whole flower.



(a)



(b)

Figure 6. Color image segmentation, (a), 4 seed points are needed when using the original live wire algorithm, (b), only 2 seed points are required when adopting the improved live wire, k=0.85.

IV. CONCLUSION

Optimal path between two seed points is redefined as the path with the minimum average cost instead of the cumulative cost. This improvement makes the "live wire" between two seed points act like a noose with adjustable elasticity. The "live wire" could reach those regions far away from the direct line between seed points by adjusting the elasticity controlling parameter. This work provides a more general version of live-wire algorithm, which is able to delineate the boundary of target object with less time demand by reducing the requirement for seed points.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of Jiangsu Province, China (Contract No: SBK20130160).

REFERENCES

[1] E.N. Mortensen, W.A. Barrett, Intelligent scissors for image composition, Proceedings of the Computer Graphics (SIGGRAPH `95), Los Angeles, CA, 1995, pp. 191-198.

[2] W.A. Barrett, E.N. Mortensen, Interactive live-wire boundary extraction, Medical Image Analysis, 1 (4) (1997) 331-341.

[3] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active Contour Models, International Journal of Computer Vision, 1 (4) (1988) 321-331.

[4] C. Xu, J.L. Prince, Snakes, shapes and gradient vector flow, IEEE Transaction on Image Processing, 7 (3) (1998) 359-369.

[5] A.A. Amini, T.E. Weymouth, R.C. Jain, Using Dynamic Programming for Solving Variational Problems in Vision, IEEE Transaction on Pattern Analysis and Machine Intelligence, 12 (2) (1990) 855-866.

[6] D. Daneels, et al., Interactive Outlining: An Improved Approach Using Active Contours, Proceedings of Storage and Retrieval for Image and Video Databases, vol. 1908, 226(1993), pp. 226-233.

[7] D.J. Williams, M. Shah, A Fast Algorithm for Active Contours and Curvature Estimation, CVGIP: Image Understanding, 55 (1) (1992) 14-26.

[8] T.F. Chan, L.A. Vese, Active Contours Without Edges, IEEE Transactions on Image Processing, 10 (2) (2001) 266-277.

[9] E.W. Dijkstra, A note on two problems in connexion with graphs, Numerical Mathematik, 1 (1959), pp. 269-270.

[10] A. Chodorowski, U. Mattsson, M. Langille, G. Hamarneh, Color Lesion Boundary Detection Using Live Wire, SPIE medical imaging, 5747 (2005) 1589-1596.

[11] The Matworks Inc. MATLAB Image Processing Toolbox User's Guide, Natick, Massachusetts, 2010, available at: http://www.mathworks.com.

[12] T.H. Hsu, C.H. Lee, L.H. Chen, An Interactive Flower Image Recognition System, Multimedia Tools And Applications, in press, 2010.

[13] J.H. Kim, R.G. Huang, S.H. Jin, K.S. Hong, Mobile-Based Flower Recognition System, Third International Symposium on Intelligent Information Technology Application, vol. 3, 2009, pp.580-583.

[14] J. Zou, G. Nagy, Evaluation of Model-Based Interactive Flower Recognition, Pattern Recognition, 2 (2004) 311-314.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Quick Capture and Reconstruction for 3D Head

Chao Lai
School of Computer
National University of Defense Technology
Changsha, China
lclaichao@163.com

Fangzhao Li
School of Computer
National University of Defense Technology
Changsha, China
lifangzhao12@163.com

Shiyao Jin
School of Computer
National University of Defense Technology
Changsha, China
syjin1937@163.com

***Abstract*—Recently, we have seen a growing trend in the reconstruction of accurate geometric models, acquired by an active method such as laser scanning or structured light. In this paper, we address the problem of creation of 3D head models from multiple RGB-Depth cameras. We propose a quick capture procedure that aims at achieving high accuracy and low-cost 3D head models with high quality. This is processed with registration of multiple RGB-Depth cameras, consolidation and matching of raw data, surface reconstruction and mesh coloring. Experiments demonstrate that it could capture the subject in one second and reconstruct the 3D head models with photo-realistic quality in 3 minutes.**

Keywords—one second capture, 3D head, high quality.

I. Introduction

A mainstream goal in computer graphics is to create quick and efficient method for building geometric models for humans. In recent several decades, as the promotion of more and more applications in the fields like video game, films and medical analysis, there is a growing trend in the reconstruction of accurate geometric models, acquired by an active method such as laser scanning or structured light. In particular, it is currently a hot topic that the reconstruction of 3D head models in academia [12,13] and industry[14]. Due to development of 3D scanning and computation, the process of reconstruction of 3D models become simply with few interactions, and the output is accurate geometric models with high quality. The underlying pipeline of all these systems is essentially the same: A person keeps static, scanned by facility, the resulting model is processed with calculation and a few interactions. However, all systems have a similar drawback in practice, which is time consuming in scanning stage, at least several minutes. Consequently, human beings, a non-rigid object, hardly keep absolute static for such a long time. Due to the slight motion of his or her body, it leads to significant distortions of the scanning result, which degrade the reconstruction quality. Thus, it is especially significant to increase capture speed.

To reduce the capture time, we propose a quick system composed of multiple RGB-Depth cameras to capture subject in one second, taking RGB images and depth data around the subject respectively without restrictions or constrains. It covers hu-man head in the full field through distributing the cameras in appropriate location and registering them within the global coordinate system. Finally, a 3D head model with high quality can be produced by further process of raw data.

II. Related Work

It is a classical topic that reconstruction of 3D head models in computer graphics, which can trace back to the parametric method of human face in the 1970s [1]. In recent years, the emergency of laser scanning facility, structured light and hybrid RGB-Depth cameras pave the way to automatic and quick 3D data acquisition of human beings. Due to the leap development of device, it leads to a comprehensive promotion of techniques and algorithms for 3D head capture and reconstruction.

For laser scanning facility, it is representative that full head scanner system [2] produced by Cyberware can reconstruct the whole body geometric model. It scans the human body from top to bottom, and calculates the surface features based on laser measuring point. Despite the high accuracy of the geometric models, achieving 0.01 mm, such methods require an expensive and cumbersome facility, priced 500,000 RMB, as well as low capture speed of 20 seconds.

For structured light scanner, InSpec Corporation proposes a capture system com-posed of several scanners, taking raw data around the subject from the scanner views respectively, and it totally covers the subject without blind spot, based on full-space-cover principle. Due to the limited view range in single frame, such method achieves capture speed of 10 seconds, priced 200,000 RMB, accuracy within 0.1 mm.

Captured Dimension Corporation proposes camera system composed of multiple high-resolution cameras in arrays, obtaining raw data by simple stacking of the initial views in a flash. With the advantage of high accuracy and flash capture, however, it is difficult to construct the complex architecture of the expensive system with so many cameras' synchronization, more than 20.

KinectFusion [3] is a low-cost RGB-Depth camera, owing to acquiring entire 3D data of subject, the subject should be scanned all around with the facility. In contrast to the previous facilities, it is convenient and low-cost. Such advantages, nevertheless, are traded for the time consuming of capture. That means, during the acquisition of data, it is unbearable for the person to keep static for a long time. For instance, it takes 3minutes to scan over the entire head. Obviously, it's rigorous and irrational.

To deal with the drawback, Tong et al. [4] propose a system composed of 3Kinect cameras and a revolving stage, which decreases the capture time of human body to 30 seconds.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.95

Fig. 1. The pipeline of our algorithm

In this paper, we consider acquisition setups consisting of a set of RGB-Depth cameras, such as Kinect. We focus on the quick capture and reconstruction of human head, that's commendable complementary for this system of human body [4].

According to the previous analysis, it suggests that scan facility is a key component of capture and reconstruction, especially conclusive impact on the accuracy of geometric model, capture time and expenses. With the precondition of high accuracy, it is inevitable to reduce capture time and expense as much as possible for the widely applicable system of capturing 3D head. For this purpose, we propose a quick capture system of 3D head, composed of 5 low-cost RGB-Depth cameras, which could capture the subject in one second and reconstruct the 3D head models with photo-realistic quality in 3 minutes.

III. Overview

Figure 1 shows a diagram of our pipeline. We propose a multi-view system composed of 5 RGB-Depth cameras around the subject, all of which are registered within the global coordinate system. Our algorithm takes as input several color images and depth data of a person's head from this multi-view system. Then a smooth and accurate geometry is obtained by applying ICP algorithm [6] to register the point clouds optimized by joint bilateral filtering and Sobel operator[7]. Next, Poisson surface reconstruction [8] is applied over the geometry surface to achieve the desirable 3D mesh. The final output of our system is a smooth mesh with photo-realistic quality through coloring the mesh.

IV. Space Configuration of Multi-View System

A. Accuracy of Camera

In this paper, we consider the low-cost RGB-Depth camera, ASUS Xtion Pro live camera, same with Kinect. Khoshelham et al. [5] analyze the accuracy and resolution of depth data and the geometric error can be written as:

$$\delta = 2.85\times10^{-6}\mathrm{d}^2 \quad (1)$$

Here, d is the distance from the center of camera to captured vertex. Thus, it is important to shrink the distance as close as possible for high accuracy.

B. Cameras Configuration

Due to the geometric features of head, similar to a sphere, we regard the head as a sphere, of which the radius is 15cm. According to this assumption, we formulate the cameras' configuration into a mathematical model. In other words, we must confirm the number of camera N and space distribution of them, with the constraint that the upper bound of error δ less than the given threshold γ as well as the whole view range R of N cameras covering the head sphere S(H). Its formulation as follows:

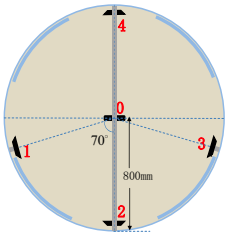


Fig. 2. Cameras configuration

$$\begin{cases} arg\min_{N}(\bigcup_{i=1}^{N} R(i) \supseteq S(H)) \\ s.t. \delta < \gamma \end{cases} \quad (2)$$

Theoretically, we just need 3 cameras that will cover the entire head. In practice, however, there is so few overlap between scanning areas of cameras that the accuracy of resulting geometry is not desirable. Moreover, it is important of face details for the quality of reconstruction. Thus, we use 5 cameras to scan the entire head. One is directed towards the face, another one towards the back, another two locate at the two sides of face to link up the facial and back data, last one covers the top(Figure 2).

C. Multi-Camera Registration

Once the multiple cameras are located, they should be registered within the global coordinate system, and, consequently, for each point of the model it is possible to match with the others, that reconstruct the entire 3D models.
In this paper, we take a head model as a static template, capture each local data of the template model from every camera respectively, and then consider Iterative Closest Point (ICP) algorithm [6] to register the multiple data and template model into the global coordinate system. The ICP algorithm is based on Least Square, and performs iteratively calculation of transformation to establish a correspondence between the point sets. Considering two point sets A and B captured by different cameras, we formulate the function to calculate the spatial transformation between them for registration as follows.

$$f(R,T) = \arg\min \sum_{i=1}^{N} \left\| B_i - (RA_i + T) \right\|^2 \quad (3)$$

Here, R is a 3D rotation matrix, and T is a 3D translation vector. It's preprocessed that the cameras registration, and once

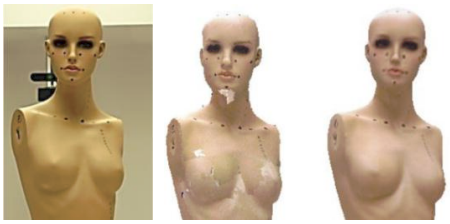


Fig. 3. The multiple cameras are registered with template model. Left is a raw data captured by one camera, middle is the registered result, and right is the final reconstruction.

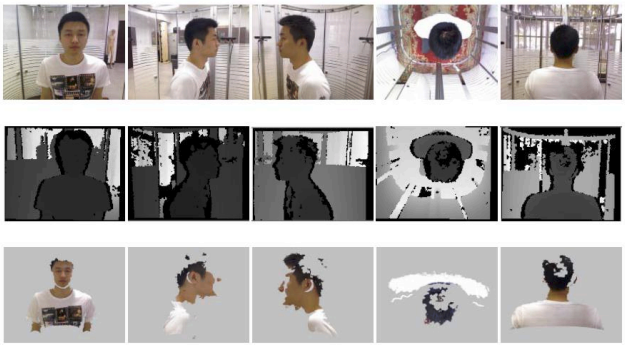


Fig. 4. Raw captured data. From top to bottom row, it is respectively RGB images, depth image and point cloud with color.

the transformation parameters are defined, it can be used repeatedly without change to match the different points set, leading to the final reconstruction (Figure 3).

V. Reconstruction of 3D Head

A. Data Capture

With the purpose of quick capture, the person would be captured in a single shot from multiple views simultaneously to cover the full head. According to values of depth images, a 3D point cloud is created with color of RGB images (Figure 4).

B. Optimization and Registration

To ensure the accuracy of face details, we set the camera, which is directed towards the face, to capture multi-frame data. As a result, it is sufficient for the face details by calculating the average of the multi-frame depth data.

Furthermore, due to the accuracy limitation of camera, joint bilateral filtering algorithm [7] is applied to smooth and optimize the noisy raw depth data, which removing out the irrelevant points, such as the points of background, leading to the remains almost belonging to the head. However, in practice, the filtered points still contain some noisy points, such as the ground and wall. Especially at the boundary, there are many points of wrong color or depth, resulting in undesirable reconstruction of model finally. As the distance from each camera to the head is equivalent, we examine the boundary by a Sobel operator [7] and set a constant threshold of depth to deal with these points. That means, all the points with depth value exceeding the threshold will be excluded, which ensure

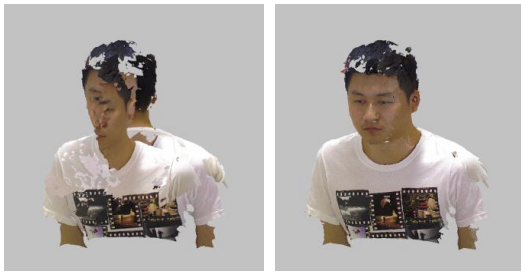


Fig. 5. Point clouds registration. Left - initial multi-view point clouds. Right - the registered point clouds

Fig. 6. The 3D head model is obtained by surface reconstruction and mesh coloring. Left is a reconstructed 3D mesh and right is the color mesh.

the remains with high accuracy. It achieves the best result while the threshold is equal to 80mm, obtained by a large variety of experiments.

According to the previous description of multi-camera registration, the primary result is obtained by making use of the parameters. Furthermore, that applying ICP algorithm [6] to the point clouds for partial calibration will effectively improve the quality of the registration. Figure 5 shows the result of registered point cloud, it appears very well with the remarkable feature of face.

C. Surface Reconstruction

Once the point clouds are registered, it is essential to reconstruct a 3D mesh by some method of surface reconstruction. Due to the advantage of robust, insensitivity to noisy data and high quality of reconstruction, a 3D mesh is obtained by apply Poisson surface reconstruction algorithm [8] to the registered point cloud (Figure 6).

D. Mesh Coloring

It is the last step of reconstruction that back-projecting color onto the 3D mesh model. Due to the different view and illumination variation across the cameras, setting the color at a vertex to the color of the nearest scan point results in visually discontinuities at adjacent transitions. Instead, by pulling color gradients [10,11] from the closest scan point, we obtain a vector field describing the local change in the texture over the mesh. Solving the Poisson equation for color field that best fits these gradients produces a seamlessly textured surface [9].

For given constraint vector field V(p), we need to obtain the coefficients of the function F whose gradients most closely match it, resulting in the function (4):

$$\arg\min_{\mathrm{F}} \int_{F} \left\| \nabla \mathrm{F} - \vec{\mathrm{V}} \right\|^2 \mathrm{dp} \qquad (4)$$

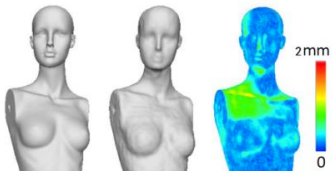


Fig. 7. Our reconstructed model is compared with the template for geometric error. Left is a template model, middle is our reconstructed model and right shows the geometric error.

Fig. 8. Here we show some reconstruction results. The upper row is the reconstructed 3D mesh, and the bottom is the colored mesh.

It can be simplified to this following equation (5).

$$Lx = b \tag{5}$$

Here, L is the Laplacian operator, $x=[x_1,\ldots,x_n]^T$ are the coefficients, and $b=[b_1,\ldots,b_n]^T$ are the integrated inner products of this vector field with the surface gradients:

$$b_i = -\int_F \left\langle \vec{V}(p), \nabla Fv_i(p) \right\rangle dp \tag{6}$$

For the resulting system of large sparse linear equations, a solution can be computed very quickly using multi-grid methods. Figure 6 shows the results of the experiment, note that even without any intelligent gradient selection, the final color seamlessly transitions across the boundaries without blending artifacts, despite the varying lighting condition across the different cameras.

VI. Result

To evaluate the quality of our solver, we compared our reconstructed 3D model with the template model obtained by laser scanning facility. Figure 7 shows the results of the experiment, with the template model on the left, our reconstructed model in the middle, and the geometric error on the right. Note that the maximum of geometric error is less than 2mm.

All the procedure of our method has been programmed in Open GL on the platform of Visual Studio 2010. The experiments are conducted on a computer equipped with an Inter Xeon E5-2643 CPU with 3.3GHz, 12GB RAM and an NVIDIA Quadro K4000 graphics card. Several results obtained by our approach are available in Figure 8. On the average, it takes 0.9 second for capture and less than 3 minutes for reconstruction of the head.

VII. Conclusion

We propose a quick capture and reconstruction system for 3D head with Depth data captured by low-cost cameras in general scenes. This is enabled by space configuration and registration of multi-view cameras that cover the entire head, optimization of raw data, surface reconstruction and mesh coloring. The results demonstrate that it could capture the subject in one second and reconstruct the 3D head models with photo-realistic quality in 3 minutes.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (grant No.61103084 and 61272334).

References

[1] Parke F I. A parametric model for human faces. [2014-05-30].http://content.lib.utah.edu/cdm/singleitem/collection/uspace/id/1613/rec/55, 2014

[2] Cyberware, Head & face color 3D scanner. [2014-05-30]. http://cyberware.com/products/scanners/px.html, 2014.

[3] Izadi S, Kim D, Hilliges O et al. Real-time 3D reconstruction and interaction using a moving depth camera// Proceedings of the 24th annual ACM symposium on User interface software and technology. New York: ACM Pess, 2011: 559-568

[4] Tong J, Zhou J, Liu L G, et al. Scanning 3d full human bodies using kinects. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(4): 643-650

[5] Khoshelham K, Elberink S O. Accuracy and resolution of kinect depth data for indoor mapping applications.Sensors,2012,12 (2): 1437-1454

[6] Besl P J, McKay N D. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 239-256

[7] Richardt C, Stoll C, Dodgson N A et al. Theobalt C. Coherent spatiotemporal filtering, up sampling and rendering of RGBZ videos. Computer Graphics Forum, 2012, 31 (2): 247-256

[8] Kazhdan M, Bolitho M. Hoppe, H. Poisson surface reconstruction.http://research.microsoft.com/enus/um/people/hoppe/proj/poissonrecon/, 2014

[9] Chuang M, Luo L J, Brown B J et al. Estimating the Laplace-Beltrami operator by restricting 3D functions. Computer Graphics Forum, 2009, 28(5): 1475-1484

[10] Perez, P., Gangnet, M., Blake, A. 2003. Poisson image editing. ACM Transactions on Graphics 22, 3, 313–318.

[11] Fattal, R., Lischinskl, D., Werman, M. 2002. Gradient domain high dy-namic range compression. ACM Transactions on Graphics 21, 3, 249–256.

[12] Li, H., Vouga, E., Gudym, A., Luo, L., Barron, J. T., Gusev, G. 3D Self-portraits. ACM Trans. Graph. 32, 6, 187:1–187:9(2013).

[13] Sturm, J., Bylow, E., Kahl, F., Cremers, D. CopyMe3D: Scanning and printing persons in 3D. In Pattern Recognition, vol. 8142. 405–414(2013).

[14] 3DSystems, http://www.3dsystems.com/.

Locality-constrained Linear Coding Based on Principal Components of Visual Vocabulary for Visual Object Categorization

Hongxia Wang, Long Zeng, Dewei Peng, Feng Geng
School of Computer Science & Technology
Wuhan University of Technology
Wuhan, Hubei, China
e-mail:whx_green@163.com; 2501526440@qq.com; 617068@qq.com; 258156120@qq.com

***Abstract*—Through the linear correlation analysis between the local feature and its K-nearest-neighbor visual words and significance testing of locality-constrained linear coding, this paper finds that the fundamental reason for causing nonsignificance of the weight coefficient is the multicollinearity of K-nearest-neighbor visual words in Locality-constrained Linear Coding (LLC) scheme. Locality-constrained principal component linear coding can solve the multicollinearity and improves the classification accuracy, but it increases the time overhead of the coding. This paper presents an improved scheme called Locality-constrained linear coding based on the principal components of visual vocabulary. To determine the principal components of K-nearest-neighbor visual words of each local feature is simplified to only determine the principal components of visual vocabulary. Experiments have been conducted for comparing and evaluating the proposed method utilizing the Caltech-4 dataset. Experimental results show that locality-constrained linear coding based on the principal components of visual vocabulary reduces the time overhead and the same time it retains the advantages of Locality-constrained principal component linear coding.**

Keywords-visual object categorization; local feature linear coding; principal components

I. INTRODUCTION

Visual object categorization is a very challenging key technology in the application of image retrieval, massive video data, visual perception alternative, automatic robot, interactive games, and etc. The bag-of-visual-words (BOV) model is the most popular, is also the most successful visual object categorization model. BOV consists of local feature extraction, visual vocabulary generation, image quantization and image classification. The image quantization uses statistical information coding technique to quantify an image.

Local feature coding is essentially allocation from local feature points to visual words, coding error is the main factor affecting the accuracy of the image classification. Vector quantization (VQ) [1-2] method is the basic method of encoding local features, for every local feature extracted from the image, finding a nearest visual word to represent it. Recently, some advanced encoding method took place of VQ, they retain more local feature information of the original image. In order to make image quantification results to be able to integrate better with linear classifier, researchers started to take up local feature linear coding method research. Locality-constrained Linear Coding (LLC) [3] has been improved on the base of linear coding methods such as sparse coding (SC) [4] and local Coordinate Coding (LCC) [5]. By far it is the local feature linear coding method getting the best visual object classification results. But, LLC method uses the least squares method to estimate the weight coefficient of nearest visual words of the local feature, it will lose some of good statistical properties when it is in the multicollinearity environment, so there exists multicollinearity problem of K-nearest-neighbor visual words and time overhead problem caused by solving multicollinearity in locality-constrained linear coding models.

This paper proposes an improved scheme called Locality-constrained linear coding based on the principal components of visual vocabulary. To determine the principal components of K-nearest- neighbor visual words of each local feature is simplified to only determine the principal components of visual vocabulary. The number of principal components of the visual vocabulary is determined according to the cumulative contribution ratio. Experimental results show that linear encoding time is reduced by 1/3 using the proposed method and the classification accuracy is improved; in the case of the comprehensive consideration of coding time and classification results, the proposed method is optimal when the cumulative contribution ratio is 85% as well as the number of principal components of the visual vocabulary is 20.

II. LINEAR CODING MODEL

A. Linear Coding Model

Local feature coding has local linear property, so one local feature point $x_i\ (i = 1, \cdots, N)$ can be linear represented by its K-nearest-neighbor visual words, building a multiple linear regression model.

$$\begin{cases} x_i = \sum_{j=1}^{K} \beta_j c_j + \varepsilon \\ \varepsilon \sim N(0,\ \sigma^2) \end{cases} \tag{1}$$

where $c_j\ (j = 1, \cdots, K)$ is nearest visual word of local feature pointer x_i, β_j is the weight value of nearest visual word c_j, ε is a unobservable random variable.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.96

For all local feature points, multiple linear regression model in matrix form is as:

$$\begin{cases} X = \boldsymbol{C}\beta + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} \sim N(0,\ \sigma^2 \boldsymbol{I}) \end{cases} \tag{2}$$

LLC method uses the least squares method to estimate the weight value of β_j from the regression model. Therefore, local feature coding model is transformed into linear coding model as formula (3) .

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left\| x_i - C_i B_i \right\|^2 \tag{3}$$

where $X = \{x_i, i = 1, \cdots, N\}$ is a local feature point set of an image, $\boldsymbol{\beta} = \{B_i, i = 1, \cdots, N\}$ is coding set of local feature points. Because a local feature point is linear represented by K-nearest-neighbor visual words, for each local feature point coding $B_i = \{\beta_j, j = 1, \cdots, K\}$, C_i is a set of K-nearest-neighbor visual words, $C_i = \{c_j, j = 1, \cdots, K\}(C_i \subset C)$, $\boldsymbol{C} = \{C_i,\ i = 1, \cdots, N\}$.

B. Problems to be solved

In [6], it is analyzed and illustrated that linear coding model have overall significant property, while some single weight coefficient is not significant. And the reason is analyzed that some nearest visual word's linear effect on local feature points is replaced by other nearest visual words. It indicates that there exists highly linear correlation among K-nearest-neighbor visual words of local feature, called multicollinearity.

The multicollinearity of K-nearest-neighbor visual words makes the performance of least squares estimation became so bad. It will bring adverse affect on two aspects: poor stability and big mean square error, and thus lead to reduction of the eventual recognition accuracy rate. In [6], use principal component estimation method to solve the problem. Firstly, calculate principal components of K-nearest-neighbor visual words. Secondly, allocate the weight coefficients of these principal components with the least squares estimation method. Then, return the weight coefficients of these principal components to the weight coefficients of K-nearest-neighbor visual words. This process of weight coefficient allocation is for each local feature. Compared to LLC method, it adds two processes: one is to determinate principal components of K-nearest-neighbor visual words; the other is the return process from the weight coefficients of these principal components to the weight coefficients of K-nearest-neighbor visual words. It increases the time overhead of local feature linear coding.

III. IMPROVED METHOD

This paper presents an approximate method. It transfers the determination of principal components of K-nearest-neighbor visual words for each local feature to the determination of principal components for the visual vocabulary. Thus, it only needs the determination of principal components once. The principal components of visual vocabulary we got are non-multicollinear. The principal components of visual vocabulary and each local feature do multiple linear regression, and adopt the least squares estimation method to get the weight coefficient of every principal component of visual vocabulary. Then return the weight coefficients of principal components of visual vocabulary to each visual word, so we can get the weight coefficient of every visual word in the visual vocabulary. After that, take the weight coefficients of K-nearest-neighbor visual words of each local feature as each local feature's final code.

A. Principal Components of Visual Vocabulary

To determine the principal components of K-nearest-neighbor visual words of each local feature is simplified to only determine the principal components of visual vocabulary once before the linear coding for all local features. Identify the principal components of visual vocabulary, as for visual vocabulary $C = \{c_i, i = 1, \cdots, n\}$, the number of the principle components is pca_n , the transformation matrix composed of top pca_n largest eigenvalues' corresponding eigenvectors is $P = (p_1, p_2, \cdots, p_{pca\_n})$. So that the principal components of visual vocabulary Y are:

$$Y = CP \tag{4}$$

where $Y = (y_1, y_2, \cdots, y_r, \cdots, y_{pca\_n})$, principle component y_r :

$$y_r = p_{r1}c_1 + p_{r2}c_2 + \cdots + p_{rn}c_n = C\ p_r \tag{5}$$

The number of principal components of visual vocabulary pca_n can be decided according to the contribution rate of the principle components. The contribution rate indicates a principle component's contribution to identify the whole data. In fact the contribution rate of every principle component is the eigenvalue of covariance matrix CC' . The first principal component corresponds to the maximum eigenvalue, so it has the maximum contribution rate. The contribution rate of the second principle component is next, and so on, the contribution rate of the last principle component is the minimum. The suitable number of principal components of visual vocabulary can be determined by the accumulative contribution rate of the principle components. In general, the accumulative contribution rate is set to above 85%.

B. Locality-constrained Linear Coding Based on Principal Components of Visual Vocabulary

After obtaining the principal components of visual vocabulary, they are used for locality-constrained linear coding.

As for the set of local feature $X = \{x_i,\ i = 1, \cdots, N\}$, the linear coding model about principal components of visual vocabulary Y is:

$$\arg\min_{\gamma} \sum_{i=1}^{N} \left\| x_i - Y\gamma_i \right\|^2 \qquad (6)$$

Locality-constrained linear coding based on principal components of visual vocabulary is illustrated in Algorithm3.1.

Algorithm3.1 Locality-constrained linear coding based on principal components of visual vocabulary
input: visual vocabulary C, principal components of visual vocabulary Y, the set of local features X, the number of nearest- neighbor visual words K
output: the set of local features' codes β

step 1 search K-nearest- neighbor visual words of each local feature:

step 1.1 calculate the Euclidean distance between each local feature and each visual word in visual vocabulary,

step 1.2 for each local feature, sort the visual words according to the distance from near to far,

step 1.3 get the top K nearest- neighbor visual words of each local feature;

step 2 get weight coefficient γ about the principal components of visual vocabulary Y for all local features according to the linear coding model(6);

step 3 using transformation matrix P, according to formula $\omega = P\gamma$,get all local features' weight coefficient ω about visual words C;

step 4 get every local feature's K nearest- neighbor visual words' weight coefficient from ω, that is the set of local features' code β.

IV. EXPERIMENTS

In all experimental results, Method1 is LLC, Method2 is locality-constrained principal component linear coding and Method3 is the proposed method termed locality-constrained linear coding based on principal components of visual vocabulary.

The experiments are executed under the operating environment of the 2.1GHz dual-core CPU and a personal computer of 2G memory; n is 1024, the number of visual words in visual vocabulary C; 128-dimension SIFT [7] is used to describe the local features; 1x1, 2x2 and 4x4 sub-regions for SPM are used; the linear SVM classifier is utilized.

A. Time Cost Experiment

20 accordion images in Caltech-101 dataset are selected for linear coding。The Comparison of coding time for three linear coding methods is as shown in Fig. 1. To compare with Method2, linear encoding time is reduced by 1/3 using Method3. Method3 is for improving Method2. It retains the advantage of Method2, and shortens coding time.

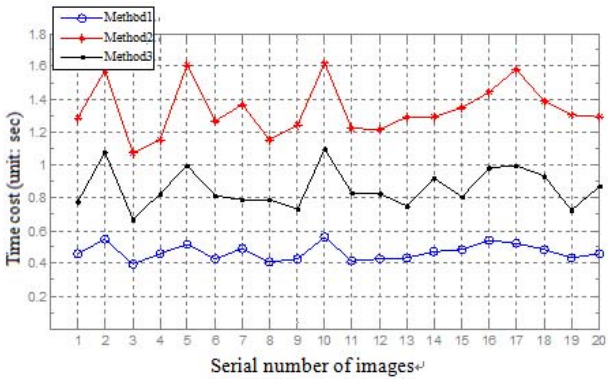


Figure 1. Comparison of time cost of three linear coding methods

B. Principal Components of Visual Vocabulary Experiment

Caltech-4 dataset is a subset Caltech-101 dataset, as shown in Fig. 2. It includes four classes of images, accordion, camera, watch and chair. It has 400 images, where each class contains the number of images ranging from 50-239. The 30 images per class have been taken out, a total of 120 images, for training the linear SVM classifier. Three methods have been run 10 times for classifying each class of objects. The classification accuracy of each class is the average value of 10 classification accuracy.

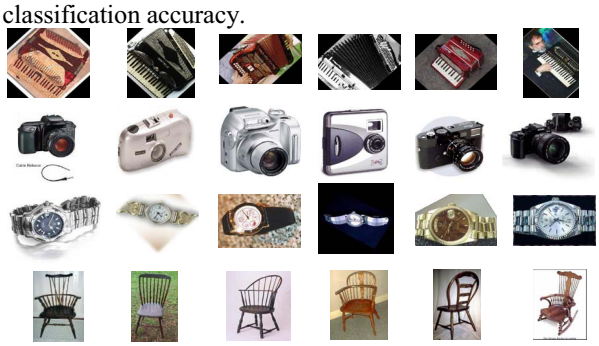


Figure 2. Caltech-4 dataset

The correspondence between number of principal components of visual vocabulary and accumulative contribution rate is as shown in Tab. 1. Classification accuracy under different accumulative contribution rate is as shown in Fig. 3. Fig. 3 shows that the average classification accuracy gradually increases when the accumulative contribution rate gradually increases from 70% to 85%; the change of the average classification accuracy is very small when the accumulative contribution rate is between 85% and 92%; the average classification accuracy begins to decline when the accumulative contribution rate increases from 92%. Moreover, if the accumulative contribution rate is higher, then the linear encoding time is longer. Considering classification accuracy and coding time, it is finally determined that the number of principal components of visual vocabulary is 20 when the accumulative contribution rate is 85% in this paper.

TABLE I. CORRESPONDENCE OF NUMBER OF PRINCIPAL COMPONENTS OF VISUAL VOCABULARY AND ACCUMULATIVE CONTRIBUTION RATE

Accumulative contribution rate(%)	*Number of Principal Components of Visual Vocabulary*
70	10
75	13
80	16
85	20
90	26
92	30
95	39

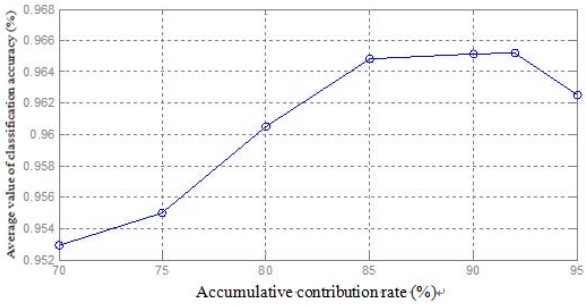


Figure 3. Classification accuracy under different accumulative contribution rate

C. The Classification Accuracy Experiment

Three methods have been run 10 times for classifying each class of objects. The classification accuracy of each class is the average value of 10 classification accuracy. In addition, the number of principal components of visual vocabulary is 20 and accumulative contribution rate is 85% in the method3.

The comparison of classification accuracy for three linear coding methods is as shown in Tab. 2. Two pooling methods have been used. It is found that the classification accuracy of the method3, Locality-constrained linear coding based on principal components of visual vocabulary, is highest.

V. CONCLUSION

LLC method has the problem of poor stability and large MSE (mean square error) because of the multicollinearity of K-nearest-neighbor visual words. The multicollinearity can be eliminated utilizing locality-constrained principal component linear coding, and classification accuracy is improved. But it has increased the time overhead of linear coding. To solve this problem, the method termed Locality-constrained Linear Coding Based on Principal Components of Visual Vocabulary is proposed in this paper. The proposed method only needs to determine the principal components of visual vocabulary once. Three experimental results show that linear encoding time is reduced by 1/3 using the proposed method; in the case of the comprehensive consideration of coding time and classification results, the proposed method is optimal when the cumulative contribution ratio is 85% as well as the number of principal components of the visual vocabulary is 20; the classification accuracy of the proposed method is best among the three methods.

ACKNOWLEDGMENT

The paper is supported by the Natural Science Funds of Hubei Province (Grant No. 2013CFB351), and the Fundamental Research Funds for the Central University (Grant No. 2014-IV-105).

REFERENCES

[1] G. Csurka, C. Bray, C. Dance, L. Fan, “Visual categorization with bags of keypoints”, ECCV Workshop on Statistical Learning in Computer Vision, pp. 1-22, 2004.

[2] S. Lazebnik, C. Schmid, J. Ponce, “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories”, Computer Vision and Pattern Recognition (CVPR), pp. 2169-2178, 2006.

[3] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, Yihong Gong, “Locality-constrained linear coding for image classification”, Computer Vision and Pattern Recognition(CVPR), pp. 3360-3367, 2010.

[4] Jianchao Yang, Kai Yu, Yihong Gong, T. Huang, “Linear spatial pyramid matching using sparse coding for image classification”, Computer Vision and Pattern Recognition (CVPR), pp. 1794-1801, 2009.

[5] K. Yu, T. Zhang, Y. Gong, “Nonlinear learning using local coordinate coding”, Proc. of NIPS’09, 2009.

[6] Ai Haojun, Zhang Min, Fang Yu, Zhao Menglei, Li Taizhou, Wang Hongxia, “Principle Component Linear Encoding for Visual Words”, The 8th Joint Confenrence on Harmonlous Human Machine Environment (HHME), 2012.

[7] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, pp. 91-110, 2004.

TABLE II. CLASSIFICATION ACCURACY COMPARISON

	Sum pooling			*Max pooling*		
	Method1	*Method2*	*Method3*	*Method1*	*Method2*	*Method3*
accordion	97.7	98.5	98.6	100	100	100
camera	77.2	82.3	81.9	95.6	95.3	95.6
watch	76.4	83.9	84.1	91.9	92.2	92.4
chair	87.5	90.7	90.2	96.1	97.7	98.0
Average value	84.7	88.9	88.7	95.9	96.3	96.5

Revealing the Structure and Function of *P. pastoris* Metabolic Network using Petri nets

Yufang WANG
dept. of mathematics and physics
jingchu university of technology
Jingmen, China
e-mail: wangyf76@126.com

Dewu DING
dept. of mathematics and computer science
chizhou college
Chizhou, China
e-mail: dwding2008@aliyun.com

***Abstract*—Due to the limitations of detailed experimental data for large scale metabolic networks, structural analysis of these networks is a growing research topic in bioinformatics field. Among several available structure-oriented approaches, Petri nets has proved to be particular efficient tool, and thus is employed for studying *P. pastoris* metabolic network in this study. After a brief introduction about the conceptual framework of Petri nets analysis, we give the first *in silico* Petri nets model of *P. pastoris* metabolic network, we also calculate, analyze and classify the p, t invariants in the model based on their structural capability and biological significance.**

Keywords-invariant analysis; metabolic model; systems biology

I. INTRODUCTION

With more and more genome-scale metabolic reconstruction available, computational methods for simulation, analysis and prediction of these models are increasingly important. Currently, lots computational modeling methods have been developed to model genome-scale metabolic networks, e.g., graph theory (or complex network theory) [1-2], constraint-based methods [3-4], and Petri nets analysis [5-6], *etc*.

Due to: 1) the conception of Petri nets is proposed to model and analyze system properties such as concurrency, communication and synchronization, which is similar to the behaviour of metabolism systems; 2) it has a visual representation that facilitates users' comprehension; 3) there is a lot of tools for modeling and simulating; and 4) it supports the integration of lots of qualitative and quantitative methods, Petri nets are particularly suitable tools for studying metabolic networks, as well as some other biological networks [5-6].

In the present article, we study *Pichia pastoris* (*P. pastoris*) metabolic network using Petri nets analysis. We firstly introduce conceptual framework of Petri nets analysis. Subsequently, we study the structural and functional characteristics of the *P. pastoris* metabolic network use the classical place/transition net, we establish the first Petri nets model of *P. pastoris* metabolic network, calculate the p, t invariants in the model, and discuss their biological significance.

II. MATERIALS AND METHODS

A. *P. pastoris Metabolic Network*

As a result of the ability to produce post-translational modifications, as well as the good protein yield/cost ratio, yeast *P. pastoris* is widely interested as a reference platform for the expression of recombinant proteins in eukaryotes [7-8]. As an important complementarity to traditional heuristical methods for the optimization of recombinant protein expression in *P. pastoris*, Tortajada *et al* recently constructed a metabolism-based model for *P. pastoris*, and performed kinds of constraint-based studies [9], their up to date metabolic network model is used in this study.

First, we get all of the chemical reaction equation involved in the model. Subsequently, in order to better reflect the conversion relationship between the substances, we have excluded current metabolites such as ADP, ATP, NADH, NADPH and H_2O. Finally, we extend the model by 6 reactions that the system interacts with the external environment, the final reactions are shown in table 1.

TABLE I. ALL REACTIONS INCLUDED IN THE MODEL OF *P. PASTORIS* METABOLIC NETWORK. IT SHOULD BE NOTED THAT ONLY PRIMARY REACTIONS AND ENVIRONMENT-RELATED REACTIONS WOULD BE INCLUDED IN OUR PETRI NETS MODEL AFTER WE EXCLUDE THE CURRENCY METABOLITES (CYT, CYTOSOLIC; MIT, MITOCHONDRIAL; E, EXTRACELLULAR).

Reaction Numbers	Metabolic Reactions	
	Reaction	***Reversibility***
Primary		
R1	GLCcyt → G6Pcyt	no
R2	G6Pcyt ↔ F6Pcyt	yes
R3	F6Pcyt ↔ FBPcyt	yes
R4	FBPcyt ↔ DHAPcyt + GAPcyt	yes
R5	DHAPcyt ↔ GAPcyt	yes
R6	GAPcyt ↔ PG3cyt	yes
R7	PG3cyt ↔ PEPcyt	yes
R8	PEPcyt ↔ PYRcyt	yes
R9	PYRcyt → OAAcyt	no

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.97

Reaction Numbers	Metabolic Reactions	
	Reaction	***Reversibility***
R10	PYRcyt → ACDcyt	no
R11	ACDcyt → ETHcyt	no
R12	ACDcyt → ACEcyt	no
R13	ACEcyt + FORcyt → ACCOAcyt	no
R14	PYRmit + FORmit → ACCOAmit	no
R15	ACCOAmit + OAAmit ↔ ICITmit + FORmit	yes
R16	ICITmit → AKGmit	no
R17	ICITmit → AKGmit	no
R18	AKGmit → SUCmit	no
R19	SUCmit → MALmit	no
R20	MALmit → OAAmit	no
R21	G6Pcyt → RU5Pcyt	no
R22	RU5Pcyt ↔ XU5Pcyt	yes
R23	RU5Pcyt ↔ R5Pcyt	yes
R24	R5Pcyt + XU5Pcyt ↔ S7Pcyt + GAPcyt	yes
R25	S7Pcyt + GAPcyt ↔ E4Pcyt + F6Pcyt	yes
R26	E4Pcyt + XU5Pcyt ↔ F6Pcyt + GAPcyt	yes
R27	DHAPcyt ↔ GOLcyt	yes
R28	OAAcyt ↔ OAAmit	yes
R29	PYRcyt → PYRmit	no
R30	AKGmit → AKGcyt	no
R31	METcyt → HCHOcyt	no
R32	HCHOcyt + XU5Pcyt ↔ DHAcyt + GAPcyt	yes
R33	DHAcyt → DHAPcyt	no
Environment-related		
gGLCcyt	GLC(E) → GLCcyt	no
gPYRcyt	PYR(E) → PYRcyt	no
gMETcyt	MET(E) → METcyt	no
gGOLcyt	GOL(E) → GOLcyt	no
g_rICITmit	CIT(E) ↔ ICITmit	yes
rETHcyt	ETHcyt → ETH(E)	no

ACCOA, acetyl-coenzyme-A; ACD, acetaldehyde; ACE, acetate; AKG, alpha-ketoglutarate; CIT, citric acid; DHA, dihydroxyacetone; DHAP, dihydroxyacetone phosphate; E4P, erytrose-4-phosphate; ETH, ethanol; F6P, fructose-6-phosphate; FBP, fructose-1,6-biphosphate; G6P, glucose-6-phosphate; GAP, glyceraldehyde-3-phosphate; GLC, glucose; GOL, glycerol; HCHO, formaldehyde; ICIT, isocitrate; MAL, malate; MET, methanol; OAA, oxaloacetate; PEP, phosphoenolpyruvate; PG3, 3-phosphoglycerate; PYR, pyruvate; R5P, ribose-5-phosphate; RU5P, ribulose-5-phosphate; S7P, septulose-7-phosphate; SUC, succinate; XU5P, xylulose-5-phosphate.

B. Petri nets Modeling

In general, the analysis of metabolic networks with Petri nets could be achieved by the classical place/transition net, which is a bipartite directed graph. It consists of two types of nodes, places $P = \{p_1, ..., p_n\}$ represent passive system elements (e.g. metabolites) and are denoted by circles in graphical representations, while transitions $T = \{t_1, ..., t_m\}$ represent active system elements (e.g. biochemical reactions) and are denoted by rectangles. Places and transitions are connected by directed arcs which are signed by corresponding stoichiometric coefficient in modeling of metabolism.

C. Invariant Analysis

Study of structural invariants which do not depend on kinetic parameters is helpful to analyze the behaviour of the system. Generally speaking, there are two types of invariants in Petri nets: p invariants (place invariants) and t invariants (transition invariants).

A p invariant is defined as a non-negative integer vector y, which holds the equation

$$y \bullet C = 0. \quad (1)$$

And a t invariant is defined as a non-negative integer vector x, which holds the equation

$$C \bullet x = 0. \quad (2)$$

Where $C(n \times m)$ is the incidence matrix of a given Petri nets, n represents the number of places and m represents the number of transitions [5].

The p invariants represent conservation relations for metabolites, while the t invariants represent groups of transitions, and the minimum t invariants correspond to elementary modes (ElMos), which are minimum sets of enzymes, that can operate together at steady state, please see ref 6 for a compare case about this point [6].

III. RESULTS AND DISCUSSION

A. The Petri nets Model

In accordance with the method described in section 2.2, using circles (place P) represent metabolites, rectangles (transition T) represent the metabolic reactions, concentric rectangular (coarse transition) represent reversible metabolic reactions, we first give all of Petri nets modeling way for 5 categories metabolic reactions in *P. pastoris* metabolic network model (see figure 1 below).

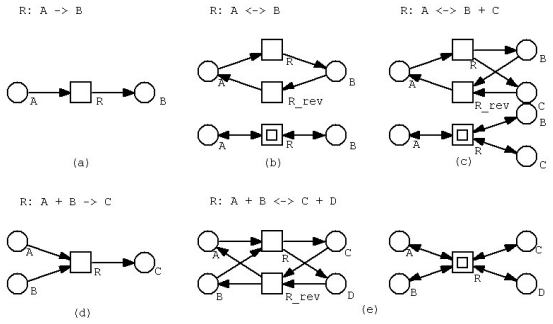


Figure 1. All 5 categories of metabolic reactions in *P. pastoris* metabolic network, and their Petri nets modeling.

Subsequently, we create the final Petri nets model for *P. pastoris* metabolic network, which consists of 35 places and 55 transitions (figure 2), among them 34 transitions (17 coarse transition) are used to represent 17 reversible reactions in *P. pastoris* metabolic network, there are: R2, R3, R4, R5, R6, R7, R8, R15, R22, R23, R24, R25, R26, R27, R28, R32 and g_rICITmit.

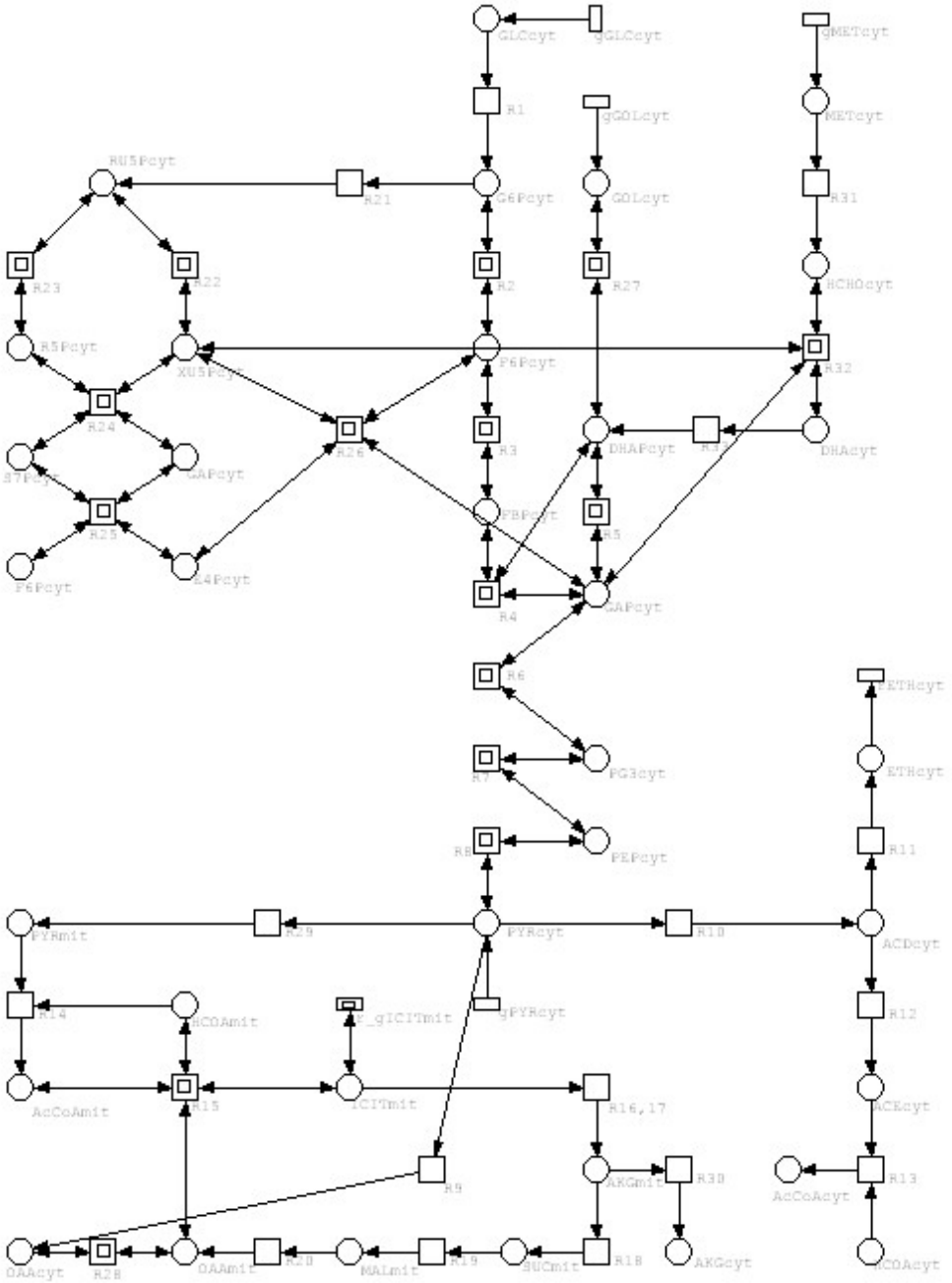


Figure 2. The Petri nets model of *P. pastoris* metabolic network.

B. Invariant Analysis

We then calculate the structural invariants (p invariants and t invariants) in the Petri nets model of *P. pastoris* metabolic network.

1) P Invariant Analysis: In general, p invariant reflects the conservation relations of the metabolites in the metabolic network models. After calculation, we get 2 group p invariants, which suggest the conservation relations between AcCoA and HCOA in the cytoplasm and mitochondria during metabolic processes.

2) T Invariant Analysis: T invariant reflects the possible reaction pathways in the metabolic network. In fact, the minimum t invariants correspond to elementary modes in

traditional biochemical pathway analysis, thus they are particularly helpful in understanding the metabolic network structure and functional characteristics. We have 30 minimum t invariants from the model, and they are classified based on their structure and functional capabilities.

a) Group 1 (17 invariants): The computation of the t invariants for the Petri nets model firstly results in 17 trivial t invariants: R2, R2_rev; R3, R3_rev; R4, R4_rev; R5, R5_rev; R6, R6_rev; R7, R7_rev; R8, R8_rev; R15, R15_rev; R22, R22_rev; R23, R23_rev; R24, R24_rev; R25, R25_rev; R26, R26_rev; R27, R27_rev; R28, R28_rev; R32, R32_rev; g_rICITmit, g_rICITmit_rev. They are all simply reversible reactions, just like type III extreme pathways in traditional biochemical pathway analysis.

b) Group 2 (1 invariant): The minimum t invariant: gMETcyt, R31, R32, R33, R4_rev, R3_rev, R2_rev, R21, R22 represents the conversion process of methanol and xylulose-5-phosphate.

c) Group 3 (4 invariants): gPYRcyt, R10, R11, rETHcyt; gGOLcyt, R27_rev, R5, R6, R7, R8, R10, R11, rETHcyt; gGLCcyt, R1, R21, R22, gMETcyt, R31, R32, R33, R5, R6, R7, R8, R10, R11, rETHcyt; and gGLCcyt, R1, R2, R3, R4, R5, R6, R7, R8, R10, R11, rETHcyt. These 4 minimum t invariants represent from different substrates (pyruvate, glycerol and glucose) to ethanol conversion process. Among them, to produce ethanol from glucose is the ethanol fermentation process (commonly known as alcoholic fermentation). These 4 minimum t invariants represent a combination of the different sub metabolic path to generates ethanol, suggest that biochemical pathways could be changed under certain conditions or requirements, reflecting the metabolism of organisms need to have a certain robustness to be used against changes in the external environment (such as mutations).

d) Group 4 (8 invariants): At last, gPYRcyt, R9, R28; gPYRcyt, R29, R14, R15, R16,17, R18, R19, R20; gGOLcyt, R27_rev, R5, R6, R7, R8, R9, R28; gGOLcyt, R27_rev, R5, R6, R7, R8, R29, R14, R15, R16,17, R18, R19, R20; gGLCcyt, R1, R2, R3, R4, R5, R6, R7, R8, R9, R28; gGLCcyt, R1, R21, R22, gMETcyt, R31, R32, R33, R5, R6, R7, R8, R9, R28; gGLCcyt, R1, R2, R3, R4, R5, R6, R7, R8, R29, R14, R15, R16,17, R18, R19, R20; and gGLCcyt, R1, R21, R22, gMETcyt, R31, R32, R33, R5, R6, R7, R8, R29, R14, R15, R16,17, R18, R19, R20, these 8 remaining minimum t invariants represents the metabolic pathway for mitochondrial oxaloacetate from different substrates (pyruvate, glycerol, and glucose), these different transformation path also confirmed the flexibility of metabolic network.

IV. CONCLUSION

Among the large number of genome-scale metabolic network modeling methods [1-6], Petri nets technology has many advantages, such as: strict mathematical description, intuitive graphical expression, numerous algorithms and analysis tools, and so on [5-6]. Furthermore, the structure of Petri nets is high corresponding to the metabolic network. Thus, Petri nets technology can easily simulate metabolic networks.

Even though *P. pastoris* metabolism has been well studied experimentally [7-9], powerful theoretical methods such as Petri nets analysis used here are still needed to discover novel mechanisms. We use the place/transition net to analyze the structural and functional characteristics of the *P. pastoris* metabolic network, we establish the Petri nets model, calculate the p, t invariants in the model, and discuss their biological significance.

ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers for their valuable comments. The work was supported by the Project of Anhui Province for Excellent Young Talents in Universities (2013SQRL096ZD).

REFERENCES

[1] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," Nat. Rev. Genet., vol. 5, Feb. 2004, pp. 101-113, doi:10.1038/nrg1272.

[2] T. Aittokallio and B. Schwikowski, "Graph-based methods for analysing networks in cell biology," Brief Bioinform., vol. 7, Jul. 2006, pp. 243-255, doi: 10.1093/bib/bbl022.

[3] N. D. Price, J. L. Reed, and B. O. Palsson, "Genome-scale models of microbial cells: evaluating the consequences of constraints," Nat. Rev. Microbiol., vol. 2, Nov. 2004, pp. 886-897, doi:10.1038/nrmicro1023.

[4] J. Schellenberger, R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, *et al*., "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0," Nat. Protoc., vol. 6, Aug. 2011, pp. 1290-1307, doi: 10.1038/nprot.2011.308.

[5] M. Herajy and M. Heiner, "Petri net-based collaborative simulation and steering of biochemical reaction networks," Fund. Inform., vol. 129, Jan. 2014, pp. 49-67, doi: 10.3233/FI-2014-960.

[6] F. Liu, M. Heiner, and M. Yang, "Modeling and analyzing biological systems using colored hierarchical Petri nets, illustrated by C. elegans vulval development," J. Biol. Syst., vol. 22, May 2014, pp. 463-493, doi: 10.1142/S0218339014500181.

[7] O. Cos, R. Ramon, J. L. Montesinos, and F. Valero, "A simple model-based control for Pichia pastoris allows a more efficient heterologous protein production bioprocess," Biotechnol. Bioeng., vol. 95, Sep. 2006, pp. 145-154, doi: 10.1002/bit.21005.

[8] M. Dragosits, J. Stadlmann, J. Albiol, K. Baumann, M. Maurer, B. Gasser, *et al*., "The effect of temperature on the proteome of recombinant Pichia pastoris," J. Proteome Res., vol. 8, Mar. 2009, pp. 1380-1392, doi: 10.1021/pr8007623.

[9] M. Tortajada, F. Llaneras, and J. Pico, "Validation of a constraint-based model of *Pichia pastoris* metabolism under data scarcity," BMC Syst. Biol., vol. 4, Aug. 2010, pp. 115, doi: 10.1186/1752-0509-4-115.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

How to Benchmark Supercomputers

Gang Xie
Institute of computer applications
China Academy of Engineering Physics
MianYang, China
e-mail: xieg@caep.cn

Yong-hao Xiao
Institute of computer applications
China Academy of Engineering Physics
MianYang, China

***Abstract*—Benchmarks are designed to mimic a particular type of workload on a component or system of supercomputer to compare platforms, identify performance bottlenecks, evaluate potential solutions, and help users in their decision of buying or using machines most appropriate for their application requirements. In this paper we discuss the purpose, significance and method of benchmarking supercomputers, describe the state of the art, and review five of the mainstream benchmarks for supercomputer evaluation, among them, Linpack is the most popular for the famous TOP500 ranking list. In addition to this, there is the HPCG benchmark the goals of which are to stress a balance of floating point operation speed and communication bandwidth and latency, there is the Rodinia benchmark for evaluating the heterogeneous supercomputers composed of both GPUs and multi-core CPUs, there is the SPEC benchmark which can be applied to the newest generation of high-performance computers, there is still the DEISA benchmark comprises a number of real applications taken from a wide range of disciplines, including astrophysics, fluid dynamics, climate modelling, biosciences, materials science, fusion power and fundamental particle physics.**

Keywords - Supercomputer; Benchmark; HPL; HPCG; SPEC; Rodinia; DEISA;

I. INTRODUCTION

To choose and buy a HPC system is not as easy as buying a PC. As HPC systems are very expensive, users usually need to evaluate a spectrum of platforms carefully and earnestly before decide which to purchase.

With the evolution of computer architectures, it has become more and more difficult to compare the performance of a variety of computer systems solely by their product specifications. Therefore we need to develop benchmarks to compare various computing platforms, position system performance bottlenecks, and evaluate design schemes. The basic requirements of a benchmark suite for general purpose computing include supporting diverse applications with various computation patterns, employing state-of-the-art algorithms, and providing input sets for testing different situations.

Benchmarks evaluate a supercomputer system by testing its operation speed, network communication, memory access and I/O etc., to reveal the strengths and weaknesses of machines of different architectures, help users in their decision of buying or using machines most appropriate for their application requirements.

Benchmarks are used to simulate a special kind of workloads on a supercomputer system, both on system components and the whole system wide. Synthetic benchmarks do this by specially created programs that impose the workload on the components, while application benchmarks impose real-world programs on the whole computer system, to test its performance. While application benchmarks usually give a much better measure of real-world performance on a given system, synthetic benchmarks are useful for testing individual components. Benchmarking is not easy and often involves several iterative rounds in order to arrive at predictable, useful conclusions. Interpretation of benchmarking data is also extraordinarily difficult.

As parallel programs complete large-scale parallel computation through inter process communication and synchronization, so the time, space and capacity characteristics of their communication mode are the major factors impacting their performance and scalability. The time characteristic of a communication mode means the frequency message is generated, while the spatial characteristic means the distribution of the message destination addresses.

II. OVERVIEW

The Linpack index of the TOP500 supercomputers show that the peak performance of high performance computers upgrades very fast. As the lifting speed of CPU performance follows the Moore's law, much faster than that of memory, hard disk and network, the IO throughput capacity gradually becomes the bottleneck of large computer systems. That's why the Graph500 benchmark was developed. At the same time, as the enhancement of network performance goes beyond people's expectations in recent years, the scale of computer systems can be expanded accordingly, often with an energy consumption of a few MWs. This causes a later operating costs surge, and concerns about power consumption. It was this concern about operating costs that gave birth to the Green500 benchmark.

Different benchmarks emphasis on different purposes. Some of them focus on the CPU performance. Some of them focus on the performance of the file server. Some of them focus on the I/O performance. Others focus on the speed of network. There are quite a few ranking lists for high performance computers. For example, the Linpack TOP500 ranking list which concerns about the CPU floating point performance, the Graph500 ranking list which concerns about the data processing ability, the Green500 ranking list which concerns about the performance and power ratio. So the evaluation of high performance computers has been controversial. But because TOP500 is very influential in the industry, it has become an honor pursued by each country in the world. This has resulted in the situation of 'big machine while small application'. The TOP500 evaluation standard needs

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.98

to be updated, to better guide the related high performance computer design work, and make it more accord with the requirement of practical applications, not just a ranking. On the other hand, in order to enhance speed and reduce energy consumption, heterogeneous computer systems consists of both GPU and multi-core CPUs are becoming more and more popular, making it necessary to establish new benchmarks to compare different architecture designs and programming environments. Thus, a new generation of evaluation tools for this class of heterogeneous systems, such as SPEC, Rodinia, and SHOC, came into being in the 2010s.[1-3]

III. FIVE OF THE MAINSTREAM BENCHMARKS FOR SUPERCOMPUTERS

There are many benchmarks of different purpose for supercomputer evaluation, some are popular and well recognized and some are less so, from among them we have chosen five mainstream ones to study. In the following subsections, we will review them one by one in turn.

A. Linpack

LINPACK uses a random number generator to generate a dense system of linear algebraic equations, and then uses partial pivoting Gauss elimination method to solve the problem. Order of the linear system of equations is determined according to 80% of the total supercomputer memory to ensure that in addition to the operating system and so on the whole coefficient matrix can be fitted into the memory system. The parallel algorithm first divides the coefficient matrix into blocks and maps the processors onto a 2D array. Then the data blocks are distributed to each node cyclically. The scalability of the algorithm is very good. No matter how large a computer system, its resources can be exhausted if the order of the linear system is set to be high enough. As the total floating point operation number of this algorithm can be figured out in advance, through dividing it by the total time consumption to complete the whole calculation we get the floating point arithmetic speed of the computer system. Because the quantity of data and calculation of this problem match well, it is quite representative as a general purpose benchmark.

LINPACK is the oldest supercomputer benchmark with rich historical data, originated in the 1990s. Because it gives a single index, it is convenient to rank supercomputers accordingly. In fact TOP500 is the most authoritative and influential ranking list for supercomputer systems in the world. Despite much of criticism and a lot of alternative benchmarking tools, its status cannot be replaced in a short time. But because Linpack's computing workload is of order three while its data size is of order two, it is not representative enough for data intensive problems. Another problem for Linpack, which can be both an advantage and disadvantage at the same time, is that its index is quite general, although convenient for ranking supercomputers, but cannot properly reflect the performance of each subsystem, such as the performance of the communication network, the memory subsystem, and the I/O server. But it is not objective to say Linpack cannot at all reflect the performance of these subsystems except CPU speed, because the computing cannot be done without the participant of all these subsystems. In fact, except huge computation amount, the memory usage and communication load of Linpack are also great. HPL, namely, the highly parallel Linpack benchmark, is proposed for modern distributed memory supercomputers. Its algorithm has such characteristics as: distributing the data blocks of the coefficient matrix cyclically onto a two-dimensional processor grid, LU decomposition with adjustable lateral communication depth, recursive panel decomposition with key search and column broadcast, bandwidth saving exchange broadcast algorithm etc.. HPL need MPI and BLAS to run. It calls the math library BLAS for solving the linear system of equations, and communicates among the nodes of the computer system through MPI. To obtain the peak CPU floating-point operation performance of HPL, we need to find the optimal combination of testing parameters such as the matrix size.

Linpack is the most popular benchmark in the world for floating point performance test of high performance computer systems. It evaluates the floating point performance of a high performance computer through solving dense linear system of algebraic equations using the Gauss elimination method. HPL is a general benchmark for modern parallel computers. Because it is simple and easy to use with a direct and reliable evaluation standard of CPU floating operation ability, it has become the fact standard of high performance computer evaluation. Both TOP500 of the world and TOPl00 of China have adopted it as the evaluation criteria to rank high performance computer systems. However, academia and industry have gotten aware of some limitations of the HPL benchmark. With high performance computer systems becoming more and more complex, it is too general to reflect the overall performance of a computer system with only one index of CPU floating point operation ability. HPL may not be able to identify possible performance bottlenecks in a computer system, thus cannot provide reliable experiences for future design and construction of supercomputer systems, Especially for multi-core clusters, the bandwidth and delay of the network subsystem, the level, access and sharing mechanism of the memory system can all restrict the performance of a supercomputer system. They can also greatly influence the results of the evaluation. So, it is not accurate enough to evaluate the performance of a computer system only with the CPU floating-point operation number.[4-6]

B. HPCG

It was widely agreed upon that, the High Performance Linpack (HPL) test is increasingly unreliable as a true measure of system performance for a growing collection of important science and engineering applications. Designing for good HPL performance can lead to design choices that are wrong for the real application mix, or add unnecessary components or complexity to the system. High performance computing applications that are governed by

differential equations, which tend to need more bandwidth and low latency and access data using irregular patterns are specifically not well served by the HPL design standards.

One factor of the supercomputer design that leads to higher speed feats under the HPL test is the GPU accelerators, which have gone popular in the supercomputing industries. These accelerators boost the performance of the top supercomputers in the Linpack tests. But the way these accelerators work doesn't always reflect real-world applications. For example, the Titan system at Oak Ridge National Laboratory was the top-ranked system in November 2012 by Linpack. However, in obtaining the HPL result on Titan, the Opteron processors played only a supporting role. All floating-point computation and all data were resident on the GPUs. In contrast, real applications will typically run solely on the CPUs and selectively offload computations to the GPU for acceleration.

On 18 November 2013, the top500 organizers released the High Performance Conjugate Gradient (HPCG) that is designed to better predict a supercomputer's real-world usefulness. The HPCG Benchmark project is an effort to create a more relevant metric for ranking HPC systems than the High Performance LINPACK (HPL) benchmark, which is currently used by the TOP500 benchmark. The goals of the new benchmark are to stress a balance of floating point and communication bandwidth and latency and to tighten the focus on messaging, memory, and parallelization. As the computational and data access patterns of HPL are no longer driving computer system designs in directions that are beneficial to many important scalable applications, HPCG is designed to exercise computational and data access patterns that more closely match a broad set of important applications, and to give incentive to computer system designers to invest in capabilities that will have impact on the collective performance of these applications.

HPCG is a complete, stand-alone code that measures the performance of basic operations in a unified code:

- Sparse matrix-vector multiplication.
- Sparse triangular solve.
- Vector updates.
- Global dot products.
- Local symmetric Gauss-Seidel smoother.
- Driven by preconditioned conjugate gradient algorithm that exercises the key kernels on a nested set of coarse grids.
- Reference implementation is written in C++ with MPI and OpenMP support.

But, according to the top500 organizers, the new HPCG test won't show real change to the list rankings too quickly, as the test will need to be run and be accepted by the supercomputing community first. "Once the definition and code for the HPCG is in a stable condition we envision collecting results for it in parallel to the ongoing effort for the HPL benchmark," said Erich Strohmaier, head of the Future Technologies Group at Lawrence Berkeley National Laboratories and Top500.org editor. "For the foreseeable future the TOP500 will be based on the HPL benchmark test but we would hope to provide additional value and information by collecting and publishing numbers for new benchmark such as HPCG as well."

As the benchmark continues to develop, we hope that it will become easier to use and optimize as well as simpler to check results within, have more relevance to a broad collection of important applications, all defined within a single number. This complementary benchmark is critical to vendors because it reflect their customer requirements. And if those demands are being met, it is vital because it means a more comprehensive approach all around to high end computing. That means scientific and enterprise benefit–it will just be a matter of time and communication to ensure this is accepted.

C. DEISA

DEISA is a benchmark suite for scientific HPC applications.

The Distributed European Infrastructure for Supercomputing Applications (DEISA) is a European Union supercomputer project funded by the European Commission. The project started in 2002 developing and supporting a pan-European distributed high performance computing infrastructure, which coupled eleven national supercomputing centers with a dedicated network connection.

DEISA produced a benchmark suite to assess the performance of parallel supercomputer systems. It provides a structured framework, which allows compilation, execution and analysis to be configured and carried out via standard input files. The benchmark comprises a number of real applications taken from a wide range of disciplines, including astrophysics, fluid dynamics, climate modelling, biosciences, materials science, fusion power and fundamental particle physics.

D. SPEC

The Standard Performance Evaluation Corporation (SPEC), founded in 1988, is an American organization aims to establish, maintain and endorse a standardized set of performance benchmarks for computers. SPEC encompasses four diverse groups: Graphics and Workstation Performance Group (GWPG), the High Performance Group (HPG), the Open Systems Group (OSG) and the Research Group (RG). The SPEC benchmarks are written in a platform neutral programming language, and the interested parties may compile the code using whatever compiler they prefer for their platform, but may not change the code. In order to use a benchmark, a license has to be purchased from SPEC. The costs vary from test to test with a typical range from several hundred to several thousand dollars. At the moment, SPEC consists of the following benchmarks:

- CPU
- Graphics/Workstations

- ACCEL/MPI/OMP

Java Client/Server
Mail Servers
Solution File Server
Power
SIP
SOA
Virtualization
Web Servers

Among them, SPEC CPU2006 tests combined performance of CPU, memory and compiler, by compute-intensive workloads. It contains two benchmark suites: CINT2006, testing integer arithmetic, with programs such as compilers, interpreters, word processors, chess programs etc., and CFP2006, testing floating point performance, with physical simulations, 3D graphics, image processing, computational chemistry etc.

The SPEC benchmarks for High Performance Computing include ACCEL, MPI2007, and OMP2012. SPEC ACCEL tests performance with a suite of computationally intensive parallel applications running under the OpenCL and OpenACC APIs. The suite exercises the performance of the host CPU, GPU, memory transfer between host and accelerator, support libraries and drivers, and compilers. MPI2007 is SPEC's benchmark suite for evaluating MPI-parallel, floating point, compute intensive performance across a wide range of cluster and SMP hardware. OMP2012 is the successor to the OMP2001, designed for measuring performance using applications based on the OpenMP 3.1 standard for shared-memory parallel processing. OMP2012 also includes an optional metric for measuring energy consumption.

E. Rodinia

The Rodinia benchmark suite is designed by University of Virginia to provide parallel programs for the study of heterogeneous systems with OpenMP, OpenCL and CUDA implementations. Rodinia 1.0 was first released on Mar 01, 2010. The newest version of Rodinia is Rodinia 3.0, released on July 23, 2014.

Heterogeneous computer systems that incorporate diverse accelerators and automatically select the best computational unit for a particular task are increasingly popular because they are becoming easier to program and offer dramatically better performance for many applications. These accelerators differ significantly from CPUs in architecture, middleware and programming models, offer parallelism at scales not currently available with other microprocessors. The performance of applications on these architectures requires taking advantage of multithreading, large number of cores, and specialized hardware. However, most of the previous benchmark suites focus on providing applications for conventional, general-purpose CPU architectures rather than heterogeneous architectures containing accelerators. They neither support these accelerators' APIs nor represent the kinds of applications and parallelism that are likely to drive development of such accelerators. Rodinia is released to address this problem. It provides publicly available implementations of applications for both GPUs and multi-core CPUs.

The Rodinia benchmark suite consists of four applications and five kernels, parallelized with OpenMP for CPUs and with the CUDA API for GPUs. The Similarity Score kernel is programmed using Mars' MapReduce API framework. Various optimization techniques and on-chip compute resources are used. The Rodinia applications cover a diverse range of domains, including Graph Algorithms, Fluid Dynamics, Physics Simulation, Pattern Recognition, Molecular Dynamics, Data Mining, etc. Each application represents a representative application from its respective domain. Users are given the flexibility to specify different input sizes for various uses.

IV. CONCLUSIONS

As the first step of supercomputer evaluation and selection, we should analyze our requirements to figure out what kind of supercomputer we want. We need to make clear the characteristics of our mainstream application programs. For example, we need to know whether our application programs are computation intensive, communication intensive, or memory intensive. Next step, we choose a few benchmarks which can well represent the characteristics of our main applications, or directly choose some of our real application programs as benchmark to evaluate our target machines. At present, Linpack is still the most recognized benchmark well representative of applications on general purpose supercomputers. On the basis of Linpack and as a complement and contrast, we should choose a few more other benchmarks to objectively evaluate our target systems.

REFERENCES

[1] Gahvari, Hormozd; Hoemmen, Mark; Demmel, James; Yelick, Katherine, "Benchmarking Sparse Matrix-Vector Multiply in Five Minutes", SPEC Benchmark Workshop, 2006.

[2] Dongarra, Jack J., "The HPC Challenge Benchmark: A Candidate for Replacing Linpack in the Top500?", SPEC Benchmark Workshop, 2007.

[3] Wong, P.; van der Wijngaart, R., "NAS Parallel Benchmarks I/O Version 2.4", NAS Technical Report NAS-03-002, NASA Ames Research Center, Moffett Field, CA, January 2003.

[4] Saphir, W.; van der Wijngaart, R.; Woo, A.; Yarrow, M., New Implementations and Results for the NAS Parallel Benchmarks 2, NASA Ames Research Center, Moffett Field, CA

[5] van der Wijngaart, R., "NAS Parallel Benchmarks Version 2.4", NAS Technical Report NAS-02-007, NASA Ames Research Center, Moffett Field, CA, October 2002.

[6] Jin, H.; Frumkin, M.; Yan, J., "The OpenMP Implementation of NAS Parallel Benchmarks and Its Performance", NAS Technical Report NAS-99-011, NASA Ames Research Center, Moffett Field, CA, October 1999.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Several stochastic gradient algorithms for nonlinear systems with hard nonlinearities

Jia Tang
Wuxi Professional College of Science and Technology
Wuxi,China
wxibm@qq.com

***Abstract*—This paper studies several identification methods for Hammerstein systems with piece-wise linearities. By using the key term separation technique, the model of the nonlinear Hammerstein systems be changed to an identification model, then based on the derived model, a stochastic gradient identification algorithm, a forgetting factor stochastic gradient algorithm and a modified stochastic gradient algorithm are used to estimate all the unknown parameters of the systems. An example is provided to show the effectiveness of the proposed algorithms.**

***Keywords*-Piece-wise linearity; modified stochastic gradient; forgetting factor; parameter estimation; nonlinear system**

I. Introduction

Nonlinear systemss are often divided into Hammerstein systems , Wiener systems and Hammerstein-Wiener systems. Hammerstein systems consist of a static nonlinear block followed by a linear dynamic block which are widely used in many areas, e.g., nonlinear filtering, actuator saturations, audio-visual processing, signal analysis. There exists a lot of work on identification for these nonlinear systems [1]–[6]. For example, the least squares (LS) algorithms [7]–[9], the stochastic gradient algorithms [10]–[12], the maximum likelihood parameter estimation algorithms [13]–[15], and the iterative algorithms [16]–[18].

The nonlinear part can be divided into the the polynomial nonlinearity or the hard nonlinearity. Some work assumed that the nonlinearity is the polynomial nonlinearity [6], [19], [20], others assumed that the nonlinearity is the hard nonlinearity [2], [3], [16], [17], [21]–[23]. Hard nonlinearity cannot be written as an analytic function of the input and is more common in engineering practice. Recently, identification of Hammerstein systems with hard nonlinearity has been received much attention [3], [17], [21], [22], [24], [25]. For example, Bai used a deterministic approach and the correlation analysis method to estimate the parameters of systems with hard input nonlinearities [21]. Chen proposed a novel estimation algorithm for dual-rate Hammerstein systems with preload nonlinearity [24], and studied identification problems for Hammerstein systems with saturation and dead-zone nonlinearities [3].

This paper deals with the identification of Hammerstein systems with piece-wise linearities. By using the key term separation technique, the model of the Hammerstein systems can be turned into an identification model, then based on the derived model, a stochastic gradient algorithm (SG)and an M-SG algorithm are proposed to estimate the unknown parameters of the systems.

Briefly, the paper is organized as follows. Section II describes the piece-wise linearities and derives an identification model. Section III studies estimation algorithms for the identification model. Section IV provides an illustrative example. Finally, concluding remarks are given in Section V.

II. The piece-wise linearities

Consider a Hammerstein system

$$A(z)y(t) = B(z)f(u(t)) + v(t), \tag{1}$$

where $y(t)$ is the system output, $u(t)$ is the system input, and $v(t)$ is a stochastic white noise with zero mean, and $A(z)$ and $B(z)$ are polynomials in the unit backward shift operator $[z^{-1}y(t) = y(t-1)]$ and

$$A(z) := 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n},$$
$$B(z) := b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3} + \cdots + b_n z^{-n}.$$

The nonlinear input $f(u(t))$ is a piece-wise linearity which is shown in Figure 1 and can be expressed as

$$f(u(t)) = \begin{cases} m_1 u(t), & u(t) \geq 0, \\ m_2 u(t), & u(t) < 0, \end{cases}$$

where m_1 and m_2 are the corresponding segment slopes.

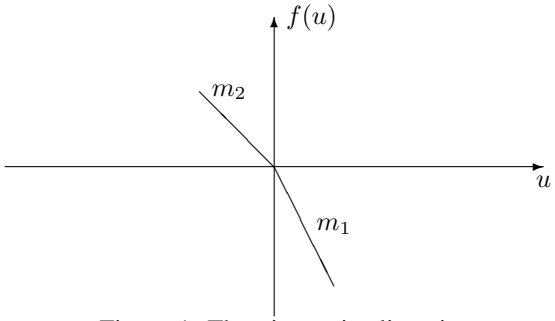


Figure 1: The piece-wise linearity

The nonlinear input $f(u(t))$ is always called the hard nonlinearity, because it cannot be written as an analytic function of the input. In order to simplify the input, define a switching function,

$$h(t) := h[u(t)] = \begin{cases} \frac{1}{2}, & u(t) \geq 0, \\ -\frac{1}{2}, & u(t) < 0. \end{cases}$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.99

Then the output $y(t)$ can be written as

$$f(u(t)) = (m_1 - m_2)u(t)h(u(t)) + \frac{1}{2}(m_1 + m_2)u(t), \quad (2)$$

and Equation (1) can be written as

$$\begin{aligned} A(z)y(t) = B(z)&((m_1 - m_2)u(t)h(u(t)) \\ &+ \frac{1}{2}(m_1 + m_2)u(t)) + v(t). \end{aligned} \quad (3)$$

From (3), we can see that the output $y(t)$ of the nonlinear block can be written as an analytic function of the input.

III. The estimation algorithms

Define the parameter vector $\boldsymbol{\theta}$ and the information vector $\boldsymbol{\varphi}(t)$ as

$$\begin{aligned} \boldsymbol{\theta} := [&b_1(m_1 - m_2), b_2(m_1 - m_2), b_3(m_1 - m_2), \cdots, \\ &b_n(m_1 - m_2), \frac{1}{2}b_1(m_1 + m_2), \frac{1}{2}b_2(m_1 + m_2), \\ &\frac{1}{2}b_3(m_1 + m_2), \cdots, \frac{1}{2}b_n(m_1 + m_2), \\ &a_1, a_2, a_3, \cdots, a_n]^{\mathrm{T}} \in \mathbb{R}^{3n}, \\ \boldsymbol{\varphi}(t) := [&u(t-1)h(t-1), u(t-2)h(t-2), \\ &u(t-3)h(t-3), \cdots, u(t-n)h(t-n), \\ &u(t-1), u(t-2), u(t-3), \cdots, \\ &u(t-n), -y(t-1), -y(t-2), \cdots, \\ &-y(t-n)]^{\mathrm{T}} \in \mathbb{R}^{3n}, \end{aligned}$$

gets

$$y(t) = \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta} + v(t). \quad (4)$$

If $\boldsymbol{\theta}$ has been estimated, none of the identification schemes can distinguish $b_i, i = 1, 2, 3, \cdots, n$ and $m_i, i = 1, 2$ from the estimated $\boldsymbol{\theta}$. Therefore, to get a unique parameterization, in this paper, we adopt the assumption that the first coefficient b_1 equals 1, i.e., $b_1 = 1$.

The parameter vector $\boldsymbol{\theta}$ and the information vector $\boldsymbol{\varphi}(t)$ be defined as

$$\begin{aligned} \boldsymbol{\theta} := [&(m_1 - m_2), b_2(m_1 - m_2), b_3(m_1 - m_2), \\ &\cdots, b_n(m_1 - m_2), \frac{1}{2}(m_1 + m_2), \\ &\frac{1}{2}b_2(m_1 + m_2), \frac{1}{2}b_3(m_1 + m_2), \\ &\cdots, \frac{1}{2}b_n(m_1 + m_2), a_1, \\ &a_2, a_3, \cdots, a_n]^{\mathrm{T}} \in \mathbb{R}^{3n}, \end{aligned} \quad (5)$$

$$\begin{aligned} \boldsymbol{\varphi}(t) := [&u(t-1)h(t-1), u(t-2)h(t-2), \\ &u(t-3)h(t-3), \cdots, u(t-n)h(t-n), \\ &u(t-1), u(t-2), u(t-3), \cdots, u(t-n), \\ &-y(t-1), -y(t-2), \cdots, \\ &-y(t-n)]^{\mathrm{T}} \in \mathbb{R}^{3n}, \end{aligned} \quad (6)$$

Using the following SG algorithm to estimate the parameter vector $\boldsymbol{\theta}$ in (5):

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}(y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1)), \quad (7)$$

$$\begin{aligned} \boldsymbol{\varphi}(t) = [&u(t-1)h(t-1), u(t-2)h(t-2), \\ &u(t-3)h(t-3), \cdots, u(t-n)h(t-n) \\ &, u(t-1), u(t-2), u(t-3), \cdots, u(t-n), \\ &-y(t-1), -y(t-2), \cdots, -y(t-n)]^{\mathrm{T}}, \end{aligned} \quad (8)$$

$$r(t) = r(t-1) + \|\boldsymbol{\varphi}(t)\|^2, \ r(0) = 1. \quad (9)$$

where $\frac{1}{r(t)}$ is the step-size and the norm of matrix X is defined by $\|X\|^2 := \mathrm{tr}[XX^{\mathrm{T}}]$.

The convergence of the SG algorithm is relatively slower compared with the recursive least squares algorithm. In order to improve the tracking performance but not to increase the computational effort of the SG algorithm, we can introduce the forgetting factor SG algorithm (the FF-SG algorithm for short) as follows:

$$\begin{aligned} \hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) &+ \frac{\boldsymbol{\varphi}(t)}{r(t)}(y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1)), \\ &\frac{1}{2} < \epsilon \leqslant 1, \end{aligned} \quad (10)$$

$$\begin{aligned} \boldsymbol{\varphi}(t) = [&u(t-1)h(t-1), u(t-2)h(t-2), \\ &u(t-3)h(t-3), \cdots, u(t-n)h(t-n), \\ &u(t-1), u(t-2), u(t-3), \cdots, u(t-n), \\ &-y(t-1), -y(t-2), \cdots, -y(t-n)]^{\mathrm{T}} \end{aligned} \quad (11)$$

$$\begin{aligned} r(t) = \lambda r(t-1) &+ \|\boldsymbol{\varphi}(t)\|^2, \\ &0 < \lambda < 1. \end{aligned} \quad (12)$$

The steps of computing the parameter estimate $\hat{\boldsymbol{\theta}}$ by the FF-SG algorithm are listed in the following:

1) Let $u(-j) = 0, y(-j) = 0, v(-j) = 0$, $j = 0, 1, 2, \cdots, n-1$, and give small positive number ε and small positive number λ.
2) Let $t = 1$, $r(0) = 1$, and $\hat{\boldsymbol{\theta}}(0) = \mathbf{1}/p_0$ with $\mathbf{1}$ being a column vector whose entries are all unity and $p_0 = 10^6$.
3) Collect the input-output data $\{u(t), y(t)\}$.
4) Form $\boldsymbol{\varphi}(t)$ by (11).
5) Compute $r(t)$ by (12).
6) Update the parameter estimation vector $\hat{\boldsymbol{\theta}}(t)$ by (10).
7) Compare $\hat{\boldsymbol{\theta}}(t)$ and $\hat{\boldsymbol{\theta}}(t-1)$: if $\frac{\|\hat{\boldsymbol{\theta}}(t) - \hat{\boldsymbol{\theta}}(t-1)\|}{\|\hat{\boldsymbol{\theta}}(t)\|} \leqslant \varepsilon$, then terminate the procedure and obtain the $\hat{\boldsymbol{\theta}}(t)$; otherwise, increase t by 1 and go to step 3.

The FF-SG algorithm can improve the convergence rate but not increase the computational effort. Unfortunately, when the estimate errors are approaching to zero, the estimates are shaking seriously [26], [27]. In order to overcome this shortcoming, we introduce the modified SG algorithm

(the M-SG algorithm for short) as follows:

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r^{\epsilon}(t)}(y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1)),$$
$$\frac{1}{2} < \epsilon \leqslant 1, \tag{13}$$
$$\boldsymbol{\varphi}(t) = [u(t-1)h(t-1), u(t-2)h(t-2),$$
$$u(t-3)h(t-3), \cdots, u(t-n)h(t-n),$$
$$u(t-1), u(t-2), u(t-3), \cdots, u(t-n),$$
$$-y(t-1), -y(t-2), \cdots, -y(t-n)]^{\mathrm{T}} \tag{14}$$
$$r(t) = r(t-1) + \|\boldsymbol{\varphi}(t)\|^2. \tag{15}$$

The steps of computing the parameter estimate $\hat{\boldsymbol{\theta}}$ by the M-SG algorithm are listed in the following:

1) Let $u(-j) = 0, y(-j) = 0, v(-j) = 0$, $j = 0, 1, 2, \cdots, n-1$, and give small positive number ε and small positive number ϵ.
2) Let $t = 1$, $r(0) = 1$, and $\hat{\boldsymbol{\theta}}(0) = \mathbf{1}/p_0$ with $\mathbf{1}$ being a column vector whose entries are all unity and $p_0 = 10^6$.
3) Collect the input-output data $\{u(t), y(t)\}$.
4) Form $\boldsymbol{\varphi}(t)$ by (14).
5) Compute $r(t)$ by (15).
6) Update the parameter estimation vector $\hat{\boldsymbol{\theta}}(t)$ by (13).
7) Compare $\hat{\boldsymbol{\theta}}(t)$ and $\hat{\boldsymbol{\theta}}(t-1)$: if $\frac{\|\hat{\boldsymbol{\theta}}(t) - \hat{\boldsymbol{\theta}}(t-1)\|}{\|\hat{\boldsymbol{\theta}}(t)\|} \leqslant \varepsilon$, then terminate the procedure and obtain the $\hat{\boldsymbol{\theta}}(t)$; otherwise, increase t by 1 and go to step 3.

IV. Example

Consider the following linear dynamic block,

$$[1 - 0.1q^{-1}]y(t) = [q^{-1} + 1.2q^{-2}]f(u(t)) + v(t),$$

the input $\{u(t)\}$ is taken as a persistent excitation signal sequence with zero mean and unit variance, and $\{v(t)\}$ is taken as a white noise sequence with zero mean and variance $\sigma^2 = 0.10^2$, the piece-wise linearity is shown in Figure 1 and with parameters: $m_1 = 1$, $m_2 = 0.8$. Then we have

$$\begin{aligned}\boldsymbol{\theta} &= [m_1 - m_2, b_2(m_1 - m_2), 0.5(m_1 + m_2),\\ &\quad 0.5b_2(m_1 + m_2), a_1]^{\mathrm{T}}\\ &= [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5,]^{\mathrm{T}}\\ &= [0.2, 0.24, 0.9, 1.08, -0.1]^{\mathrm{T}},\\ \boldsymbol{\varphi}(t) &= [h(u(t-1))u(t-1), h(u(t-2))u(t-2),\\ &\quad u(t-1), u(t-2), -y(t-1)]^{\mathrm{T}}.\end{aligned}$$

Applying the proposed SG, FF-SG and M-SG algorithms to estimate the parameters of this system, the parameter estimates and their errors are shown in Tables I-V.

Let $\hat{\alpha}_i$ be the ith element of the vector $\hat{\boldsymbol{\theta}}$. From the definition of $\boldsymbol{\theta}$, we have: $\hat{a}_1 = \hat{\alpha}_5$, $\hat{b}_2 = \frac{\hat{\alpha}_2}{\hat{\alpha}_1}$. Furthermore, we can compute the estimates $\hat{m}_1 = \hat{\alpha}_3 + \frac{\hat{\alpha}_1}{2}$, $\hat{m}_2 = \hat{\alpha}_3 - \frac{\hat{\alpha}_1}{2}$.

From Tables I-V, we can conclude.

Table I The SG estimates and errors

t	α_1	α_2	α_3	α_4	α_5	δ (%)
100	-0.0422	0.0043	0.4938	0.5448	-0.2357	52.9384
200	-0.0291	0.0224	0.5536	0.6047	-0.2483	47.3742
300	-0.0180	0.0347	0.5844	0.6351	-0.2537	44.4015
500	-0.0108	0.0442	0.6168	0.6669	-0.2658	41.6292
1000	0.0009	0.0534	0.6589	0.7046	-0.2646	37.9839
1500	0.0055	0.0585	0.6771	0.7217	-0.2663	36.4216
2000	0.0102	0.0630	0.6906	0.7342	-0.2663	35.2145
2500	0.0135	0.0666	0.7024	0.7442	-0.2663	34.2374
3000	0.0160	0.0690	0.7093	0.7510	-0.2653	33.5802
True values	0.2000	0.2400	0.9000	1.0800	-0.1000	

Table II The FF-SG estimates and errors($\lambda = 0.9$)

t	α_1	α_2	α_3	α_4	α_5	δ (%)
100	0.1056	0.0186	0.6563	0.6136	-0.2476	30.1432
200	0.1521	0.1031	0.7538	0.7822	-0.2194	20.0021
300	0.1713	0.1634	0.8158	0.8423	-0.1722	9.9921
500	0.1899	0.1911	0.9076	0.8672	-0.1632	5.0129
1000	0.2132	0.2048	0.9135	1.1762	-0.1306	3.0431
1500	0.2104	0.2197	0.9178	1.1347	-0.1023	2.4311
2000	0.2021	0.2356	0.9071	1.0075	-0.1011	1.9854
2500	0.1923	0.2376	0.8912	1.0123	-0.0901	1.6567
3000	0.1952	0.2421	0.9025	1.0704	-0.1027	1.0231
True values	0.2000	0.2400	0.9000	1.0800	-0.1000	

Table III The FF-SG estimates and errors($\lambda = 0.8$)

t	α_1	α_2	α_3	α_4	α_5	δ (%)
100	0.1174	0.0619	0.8128	0.9035	-0.1812	20.0681
200	0.1927	0.1290	0.9055	1.0303	-0.1449	9.0010
300	0.2224	0.1689	0.8981	1.0440	-0.1097	5.7755
500	0.2285	0.1994	0.9009	1.0734	-0.1009	3.4647
1000	0.2206	0.2145	0.9003	1.0762	-0.1026	2.2923
1500	0.2105	0.2298	0.8962	1.0786	-0.1004	1.0502
2000	0.2033	0.2333	0.8983	1.0778	-0.1023	0.5736
2500	0.1952	0.2395	0.8952	1.0825	-0.0981	0.5179
3000	0.1980	0.2361	0.8958	1.0754	-0.1055	0.6523
True values	0.2000	0.2400	0.9000	1.0800	-0.1000	

Table IV The M-SG estimates and errors($\epsilon = 0.9$)

t	α_1	α_2	α_3	α_4	α_5	δ (%)
100	0.1074	0.0193	0.6642	0.6363	-0.2334	28.0681
200	0.1624	0.1070	0.7426	0.7925	-0.2035	17.5012
300	0.1724	0.1646	0.8427	0.8545	-0.1626	8.9954
500	0.1902	0.1934	0.9114	0.9732	-0.1412	4.2514
1000	0.2132	0.2048	0.9135	1.1762	-0.1306	3.0431
1500	0.2097	0.2231	0.9154	1.1279	-0.1015	2.0002
2000	0.2036	0.2324	0.9057	1.0042	-0.1014	1.7736
2500	0.1958	0.2338	0.8984	1.0869	-0.0902	1.4567
3000	0.1964	0.2414	0.8990	1.0721	-0.1033	0.9981
True values	0.2000	0.2400	0.9000	1.0800	-0.1000	

Table V The M-SG estimates and errors($\epsilon = 0.8$)

t	α_1	α_2	α_3	α_4	α_5	δ (%)
100	0.1274	0.0693	0.8642	0.9363	-0.1734	19.0681
200	0.2027	0.1470	0.9032	1.0321	-0.1395	8.5012
300	0.2324	0.1791	0.8964	1.0542	-0.1126	4.9954
500	0.2280	0.1973	0.9014	1.0725	-0.1012	3.3528
1000	0.2147	0.2134	0.9007	1.0777	-0.1008	2.2431
1500	0.2064	0.2301	0.8973	1.0779	-0.1003	1.0002
2000	0.2019	0.2367	0.8963	1.0018	-0.1003	0.6736
2500	0.1972	0.2364	0.8984	1.0836	-0.0952	0.5567
3000	0.1982	0.2402	0.8992	1.0734	-0.1052	0.5487
True values	0.2000	0.2400	0.9000	1.0800	-0.1000	

1) The parameter estimation errors become smaller and smaller and go to zero with t increasing.

2) The FF-SG algorithm has a faster convergence rate than the SG algorithm.
3) The M-SG algorithm has a faster convergence rate than the SG algorithm.
4) When the λ is smaller, the convergence rate of the FF-SG algorithm is faster.
5) When the convergence index of the M-SG algorithm is smaller, the convergence rate is faster.

V. Conclusion

Several stochastic gradient algorithms are presented to identify Hammerstein systems with piece-wise linearity in this paper. The model of the nonlinear system be turned into an identification model by using the key term separation technique, then based on the identification model, we proposed an SG algorithm, an FF-SG algorithm and an M-SG algorithm to estimate all the parameters of the system. The simulation results verify the proposed algorithm.

References

[1] Y. Liu and E.W. Bai, "Iterative identification of Hammerstein systems," *Automatica*, vol. 43, no. 2, pp. 346-354, 2007.

[2] F. Ding, P.X. Liu, and G. Liu, "Identification methods for Hammerstein nonlinear systems," *Digital Signal Processing*, vol. 21, no. 2, pp. 215-238, 2011.

[3] J. Chen, X.P. Wang, and R.F. Ding, "Gradient based estimation algorithm for Hammerstein systems with saturation and dead-zone nonlinearities," *Applied Mathematical Modelling*, vol. 36, no. 1, pp. 238-243, 2012.

[4] L. Yu, J.B. Zhang, Y.W. Liao, and J. Ding, "Parameter estimation error bounds for Hammerstein finite impulsive response models, Applied Mathematics and Computation 202 (2) (2008) 472-480.

[5] D.Q. Wang, Y.Y. Chu, and F. Ding, "Auxiliary model-based RELS and MI-ELS algorithms for Hammerstein OEMA systems," *Computers & Mathematics with Applications*, vol. 59, no. 9, pp. 3092-3098, 2010.

[6] J. Chen, Y. Zhang, and R.F. Ding, "Auxiliary model based multi-innovation algorithms for multivariable nonlinear systems," *Mathematical and Computer Modelling*, vol. 52, no. 9-10, pp. 1428-1434, 2010.

[7] C. Wang and T. Tang, "Recursive least squares estimation algorithm applied to a class of linear-in-parameters output error moving average systems," *Applied Mathematics Letters*, vol. 29, pp. 36-41, 2014.

[8] F. Ding, "Combined state and least squares parameter estimation algorithms for dynamic systems," *Applied Mathematical Modelling*, vol. 38, no. 1, pp. 403-412, 2014.

[9] Y.B. Hu, "Iterative and recursive least squares estimation algorithms for moving average systems," *Simulation Modelling Practice and Theory*, vol. 34, pp. 12-19, 2013.

[10] Y.B. Hu, B.L. Liu, and Q. Zhou, "A multi-innovation generalized extended stochastic gradient algorithm for output nonlinear autoregressive moving average systems," *Applied Mathematics and Computation*, vol. 247, pp. 218-224, 2014.

[11] D. Valiente, A. Gil, L. Fernandez, et al., "A modified stochastic gradient descent algorithm for view-based SLAM using omnidirectional images," *Information Sciences*, vol. 279 pp. 326-337, 2014.

[12] F. Ding, "Hierarchical multi-innovation stochastic gradient algorithm for Hammerstein nonlinear system modeling," *Applied Mathematical Modelling*, vol. 37, no. 4, pp. 1694-1704, 2013.

[13] H. Karimi and K.B. McAuley, "A maximum-likelihood method for estimating parameters, stochastic disturbance intensities and measurement noise variances in nonlinear dynamic models with process disturbances," *Computers & Chemical Engineering*, vol. 67, pp. 178-198, 2014.

[14] T. Denoeux, "Maximum likelihood estimation from fuzzy data using the EM algorithm, Fuzzy Sets and Systems 183 (1) (2011) 72-91.

[15] A. M. Manceur, P. Dutilleul, "Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion," *Journal of Computational and Applied Mathematics*, vol. 239, pp. 37-49, 2013.

[16] J. Vörös, "Modeling and parameter identification of systems with multi-segment piecewise-linear Characteristics," *IEEE Transactions on Automatic control*, vol. 47, no. 1, pp. 184-188, 2002.

[17] J. Vörös, "Modeling and identification of systems with backlash," *Automatica*, vol. 46, no. 2, pp. 369-374, 2010.

[18] F. Ding, K.P. Deng, and X.M. Liu, "Decomposition based Newton iterative identification method for a Hammerstein nonlinear FIR system with ARMA noise," *Circuits, Systems and Signal Processing*, vol. 33, no. 9, pp. 2881-2893, 2014.

[19] D.Q. Wang, Y.Y. Chu, and F. Ding, "Auxiliary model-based RELS and MI-ELS algorithms for Hammerstein OEMA systems," *Computers & Mathematics with Applications*, vol. 59, no. 9, pp. 3092-3098, 2010.

[20] F. Ding, Y. Shi, and T. Chen, "Auxiliary model-based least-squares identification methods for Hammerstein output-error systems," *Systems & Control Letters*, vol. 56, no. 5, pp. 373-380, 2007.

[21] E.W. Bai, "Identification of linear systems with hard input nonlinearities of known structure," *Automatica*, vol. 38, no. 5, pp. 853-860, 2002.

[22] J. Vörös, "Parameter identification of discontinuous Hammerstein systems," *Automatica*, vol. 33, no. 6, pp. 1141-1146, 1997.

[23] J. Vörös, "Parameter identification of Wiener systems with multisegment piecewise-linear nonlinearities," *Systems & Control Letters*, vol. 56, no. 2, pp. 99-105, 2007.

[24] J. Chen, L.X. Lv, and F. Ding, "Parameter estimation for dual-rate sampled data systems with preload nonlinearities," *Advances in Intelligent and Soft Computing 2011 3rd International Asia Conference on Informatics in Control, Automation and Robotics*, pp. 43-50m, 2011.

[25] Y. Rochdi, F. Giri, J.B. Gning, and F.Z. Chaoui, "Identification of block-oriented systems in the presence of nonparametric input nonlinearities of switch and backlash types," *Automatica*, vol.46, no. 5, pp. 864-877, 2010.

[26] F. Ding, "System Identification—New Theory and Methods," *Science Press*, Beijing, 2013.

[27] F. Ding, "System Identification—Performances Analysis for Identification Methods," *Science Press*, Beijing, 2014.

Safety Assessment Model Based on Dynamic Bayesian Network

Yu Feng[1], Liu Wei[1], Gao Chunyang[1], Tan Lisha[2]
1. Network Center of Shenyang Jianzhu University, Shenyang, Liaoning, 110168, China
2. Student Work Department of Shenyang Jianzhu University, Shenyang, Liaoning, 110168, China
Email: wind@sjzu.edu.cn

***Abstract*—To take full advantage of the historical information, we adopt dynamic Bayesian networks and discrete-time Bayesian networks to describe the event trees or dynamic fault trees. On this basis of the regulation, corresponding dynamic Bayesian networks and discrete-time Bayesian networks models are established for safety assessment in timing system. When the discrete-time time Bayesian network of dynamic fault trees is acquired, the Bayesian network is used to study the consequences probability, key degree and computation method. A simulated banking transaction system is analyzed and assessed by above models. The results show that the safety assessment scheme can provide model without subjectivity, which raises the credibility of model. In addition, it can obtain general assessment results and acquire more extensive diagnostics and information.**

Keywords- safety assessment; historical information; key degree; Bayesian network; fault tree

I. INTRODUCTION

The Bayesian network technologies developed recently are used to express and analyze events in uncertainty. From the inference mechanism and state description, it has some approximates in event tree and fault tree. It has the ability to describe event polymorphism and non-determinism in logic relationship, which is very suitable for safety analysis [1]. Reference [2-4] discuss the method to transform from fault tree to Bayesian network and provide the method to transform from "AND gate", "OR gate" and "voter gate" to Bayesian networks. The transformed Bayesian networks have presented the method to calculate the probability of top event. Zhou et. al [5] study the transformation from other logic gates to Bayesian networks in fault tree and they provide solutions of the minimal path set, the minimal cut set and the importance. Chung, et. al [6] use polymorphic logic diagram to give a polymorphic reliability analysis based on Bayesian network. Professor Dugan in Virginia University who leads a research team combines Markov theory and combinatorial mathematics to establish the dynamic fault tree model. This research team provides the probabilistic safety assessment method of timing system which combines the event trees and dynamic fault trees [7]. Meanwhile, they can effectively operate the problems of sequence failure. Besides, dynamic Bayesian network has extremely mature algorithm as well as tool and software support. Discrete-time Bayesian network can adopt static Bayesian network model for calculation in quantity. In order to make full use of the historic information, this paper applies dynamic Bayesian network and discrete-time Bayesian network to describe event tree and dynamic fault tree.

II. RELATED WORKS

A. Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBN) is a graph structure established on static Bayesian network and Markov model. It is some extension of initial network on time. It consists of initial networks and transmitting networks and each time fragment corresponds to a static Bayesian network. There are finite time fragments in the whole network ($N>1$). Each fragment is composed of a directed acyclic graph $G_T =< V_T, E_T >$ and condition probability distribution satisfying the conditional independence assumption. V_T and E_T are the node set of time fragment T and directed sides. The fragments are connected by these directed sides which are called as transmitting network. E_T^{tmp} denotes the transmitting network of fragment t and T_0 denotes the initial time fragment.

$$E_T^{tmp} = \{(a,b) \mid a \in V_{T-1}, b \in V_T\}, T_0 < T \le T_0 + N\Delta T \quad (1)$$

DBN satisfies the first-order Markov approximation: The state of time fragment T is only related to the state of fragment $T-\Delta T$, and is unrelated to the state of fragments before $T-\Delta T$. That is

$$P(G_T \mid G_{T-\Delta T}, ..., G_{T_0}) = P(G_T \mid G_{T-\Delta T}) \quad (2)$$

Then

$$\begin{cases} V = V(T_0, N) = \bigcup_{T=T_0+N\Delta T}^{T_0+N\Delta T} V_T \\ E = E(T_0, N) = E_{T_0} \cup \bigcup_{T=T_0+N\Delta T}^{T_0+N\Delta T} E_T^* \end{cases} \quad (3)$$

$$E_T \in V_T \otimes V_T, E_T^* = E_T \cup E_T^{tmp} \quad (4)$$

$\otimes$ is the Cartesian product operation among the sets.

B. Key Degree Redefining of Timing System

The quantitative calculations of traditional event trees or dynamic fault trees need to be transformed into Markov chain first. Then differential equations corresponding to different chain length are to be solved.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.100

There exists defect in combinatorial explosion and it is hard to analyze large systems. While DBN only need to define the initial network and transmitting network to avoid that problem effectively. According to adjusted DBN we can use existing multiple mature algorithms to easily calculate the occurrence probability of each consequence at any time.

$$P_j(t) = P(Outcome_t = j \mid E_{01} = E_{02} \ldots = E_{0m} = 0) \quad (5)$$

$j \in o$, o is the state space of leaf node $Outcome$. E_{oj} denotes the state of root node E_i at initial hour t_0. m is the amount of root nodes in initial network.

Besides traditional analysis results, we can obtain more rich information with DBN. If some event E_m occurs at time t, the posteriori probability of event E_k occurs at each hour:

$$P(E_{tk} = 1 \mid E_{Tm} = 1) = \frac{P(E_{tk} = 1, E_{Tm} = 1)}{P(E_{Tm} = 1)} \quad (6)$$

III. SAFETY ASSESSMENT MOELING

A. Transformation of Static Logic Gates

If the task schedule of system T is divide into n intervals and the interval length within the time $\Delta = T/n$. Assume $X = [(x-1)\Delta, x\Delta]$ denotes that X loses effectiveness in $[(x-1)\Delta, x\Delta]$ ($0 \le x \le n$). $X = [T, \infty]$ denotes that X does not lose effectiveness in T and $f_X(t)$ is the failure probability density function of X. If X is a root node then its priori probability can be computed as

$$P(X = [(x-1)\Delta, x\Delta])) = \int_{(x-1)\Delta}^{x\Delta} f_X(t)dt,\ 0 < x \le n \quad (7)$$

$$P(X = [T, \infty]) = \int_T^{\infty} f_X(t)dt$$

(a) AND gate

Figure 1 describes Bayesian network corresponding to logic relation "AND". The condition probability distribution of node AND is:

$$\begin{aligned} P_{x,y,z} &= P(AND = [(z-1)\Delta, z\Delta)] \mid A) \\ &= [(x-1)\Delta, x\Delta], B = [(y-1)\Delta, y\Delta]) \quad (8) \\ &= \begin{cases} 1 & (0 < z <= \max\{x, y\} \le n) \\ 0 & else \end{cases} \end{aligned}$$

$$\begin{aligned} P_{x,\infty,\infty} &= P(AND = [(T, \infty) \mid A \\ &= [(x-1)\Delta, x\Delta), B = [(y-1)\Delta, y\Delta])) = 1 \end{aligned} \quad (9)$$

$$\begin{aligned} P_{\infty,y,\infty} &= P(AND = [(T, \infty) \mid A \\ &= [T, \infty), B = [T, \infty)) = 1 \end{aligned} \quad (10)$$

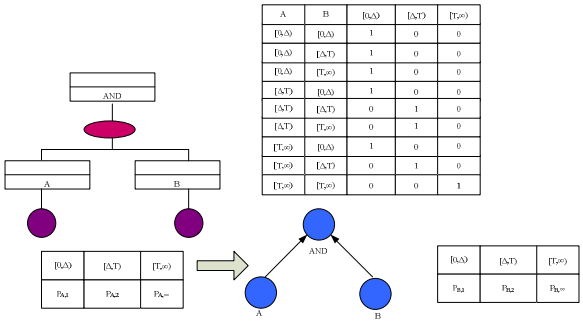


Figure 1. Discrete-time Bayesian network corresponding to "AND".

(b) OR gate

The Bayesian network corresponding to logic relation "OR" is shown as figure 2. The condition probability distribution of node OR is:

$$\begin{aligned} P_{x,y,z} &= P(OR = [(z-1)\Delta, z\Delta)] \mid A \\ &= [(x-1)\Delta, x\Delta], B = [(y-1)\Delta, y\Delta]) \quad (11) \\ &= \begin{cases} 1 & (0 < z <= \min\{x, y\} \le n) \\ 0 & else \end{cases} \end{aligned}$$

$$\begin{aligned} P_{x,\infty,x} &= P(OR = [(x-1)\Delta, x\Delta) \mid A \\ &= [(x-1)\Delta, x\Delta), B = [T, \infty])) = 1 \end{aligned} \quad (12)$$

$$\begin{aligned} P_{\infty,y,y} &= P(OR = [(y-1)\Delta, y\Delta] \mid A \\ &= [T, \infty), B = [(y-1)\Delta, y\Delta)) = 1 \end{aligned} \quad (13)$$

$$\begin{aligned} P_{\infty,\infty,\infty} &= P(OR = [T, \infty) \mid A \\ &= [T, \infty), B = [T, \infty)) = 1 \end{aligned} \quad (14)$$

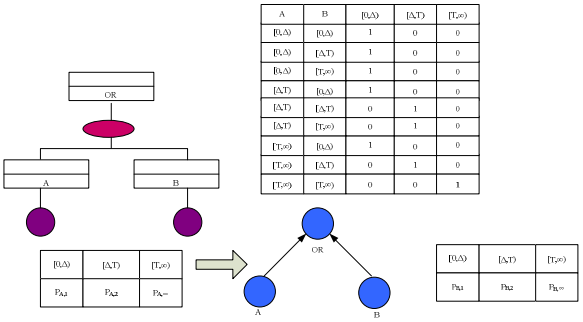


Figure 2. Discrete-time Bayesian network corresponding to "OR".

B. Transformation of Dynamic Fault Trees

(a) Sequence dependent gate

Figure 3 provides the corresponding discrete-time Bayesian network of sequence dependent gate. The condition probability distribution of node B and SEQ are described as 21-25.

$$P_{x,y} = P(B=[(y-1)\Delta, y\Delta] \mid A=[(x-1)\Delta, x\Delta))$$

$$= \begin{cases} 0 & (0 < y \le x \le n) \\ \dfrac{\int_{(y-1)\Delta}^{y\Delta}\int_{\alpha}^{y\Delta} \lambda_b e^{-\lambda_b(b-a)}\lambda_\alpha e^{-\lambda_\alpha a} dbda}{\int_{(y-1)\Delta}^{y\Delta} \lambda_\alpha e^{-\lambda_\alpha a} da} & (0 < y = x \le n) \\ = 1 - \dfrac{\lambda_\alpha (e^{(\lambda_a-\lambda_b)\Delta}-1)}{(\lambda_a-\lambda_b)(e^{\lambda_a\Delta}-1)} & \\ \dfrac{\int_{(x-1)\Delta}^{x\Delta}\int_{(y-1)\Delta}^{y\Delta} \lambda_b e^{-\lambda_b(b-a)}\lambda_\alpha e^{-\lambda_\alpha a} dbda}{\int_{(x-1)\Delta}^{x\Delta} \lambda_\alpha e^{-\lambda_\alpha a} da} & \\ = \dfrac{\lambda_\alpha e^{-(\lambda_a-\lambda_b)(y-x)\Delta}(e^{(\lambda_a-\lambda_b)\Delta}-1)(e^{\lambda_b\Delta}-1)}{(\lambda_a-\lambda_b)(e^{\lambda_a\Delta}-1)} & (0 < x < y \le n) \end{cases} \quad (15)$$

$$P_{x,\infty} = P(B=[T,\infty] \mid A=[(x-1)\Delta, x\Delta)) = 1 - \sum_{0<y\le n} P_{x,y} \quad (16)$$

$$P_{\infty,\infty} = P(SEQ=[(y-1)\Delta, y\Delta)] \mid B=[T,\infty)) = 1 \quad (17)$$

$$P_{y,y} = P(SEQ=[(y-1)\Delta, y\Delta)] \mid B=[(y-1)\Delta, y\Delta)) = 1 \quad (18)$$

$$P_{\infty,\infty} = P(SEQ=[T,\infty] \mid B=[T,\infty] = 1 \quad (19)$$

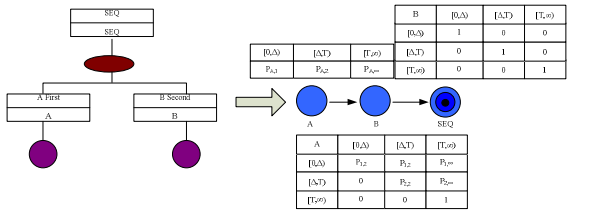


Figure 3. Discrete-time Bayesian network corresponding to sequence dependent gate.

(b) Function correlation gate

Figure 4 describes the topology structure of Bayesian network corresponding to the function correlation gate. The condition probability distribution of node A and $FDEP$ are described as formula 15-19.

$$P_{\infty,y} = P(A=[(y-1)\Delta, y\Delta \mid Tr = \int_{(y-1)\Delta}^{y\Delta} \lambda_A e^{-\lambda_A a} da = e^{-\lambda y\Delta}(e^{\lambda\Delta}-1) \quad (20)$$

$$P_{\infty,\infty} = P(A=[T,\infty) \mid Tr=[T,\infty)) = 1 - \sum_{0<y\le n} P_{\infty,y} \quad (21)$$

$$P_{x,x} = P(FDEP=[(x-1)\Delta, x\Delta \mid Tr=[(x-1)\Delta, x\Delta)) = 1 \quad (22)$$

$$P_{\infty,\infty} = P(FDEP=[T,\infty) \mid Tr=[T,\infty)) = 1 \quad (23)$$

$$P_{x,y} = P(A=[(y-1)\Delta, y\Delta \mid Tr=[T,\infty)) = [(x-1)\Delta, x\Delta))$$

$$= \begin{cases} 0 & (0 < y < x \le n) \\ \int_{(y-1)\Delta}^{y\Delta} \lambda_A e^{-\lambda_A a} da = e^{-\lambda_A y\Delta}(e^{\lambda\Delta}-1) & (0 < y = x \le n) \\ 1 - \sum_{y<x} P_{x,y} & (0 < x < y \le n) \end{cases} \quad (24)$$

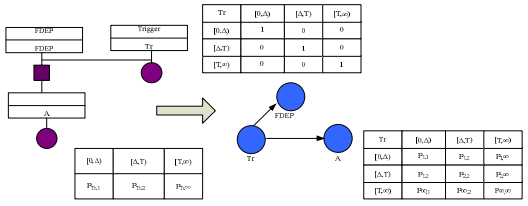


Figure 4. Discrete-time Bayesian network corresponding to function correlation gate.

C. *Transformation Algorithm for Event Trees*

(a) According to system task T, establish an $n+1$ state node in the discrete Bayesian network for initial events and link events

(b) If the amount of results is m, then establish an $m(n+1)$ state node $Outcome$

(c) Connect the nodes corresponding to initial events and link events to node $Outcome$

(d) Determine the condition probability distribution of $Outcome$ according to the follows:

$$P(Outcome = j_{i_1}\cdots_{ik}[(o-1)\Delta, o\Delta) \mid E_{i_1} = [(e_{i_1}-1)\Delta, e_{i_1}\Delta), ..., E_{i_k} = [(e_{i_k}-1)\Delta, e_{i_k}\Delta), E_{p_1} = ...E_{pl} = [T,\infty)) = 1 \quad (25)$$

$Outcome = j_{i_1}\cdots_{ik}[(o-1)\Delta, o\Delta)$ denotes the result $j_{i_1}\cdots_{ik}$ is in $(o-1)\Delta, o\Delta)$, $0 < o = \max\{e_{i_1}, ...e_{i_k}\} \le n$. $j_{i_1}\cdots_{ik}$ denotes the occurrence of link events $E_{i_1}, ..., E_{i_k}$ without other events. $k+l-1$ denotes the amount of link events.

IV. CASE STUDY

A. *Systematic Modeling for Banking Transaction System*

Banking transaction system is the transaction platform between customers and bank, or between customers themselves. Once a problem appears, it will result in disastrous consequences. The transaction system safety involves many complex factors including hardware, software, human factors etc. Now we will perform safety assessment on a simplified banking transaction system in order to illustrate the effectiveness of our method.

Generally speaking, the customers have three ways to finish the transaction through bank:

$X1$: Customers can use personal computer at home to trade through internet banking.

$X2$: At banking offices, customers can operate through filling forms by cashiers on a random client terminal.

$X3$: Banking system managers can directly finish the transactions on servers.

Based on above conditions, the event tree can be established and there are totally four consequences: OK, $F1$, $F2$ and $F3$. OK refers to normal state, that is, the transaction system is normally finished and customers successfully finish transaction. $F1$ and $F3$ refer to failure states in different degrees (transaction system is in partial failure while customers successfully finish transaction). $F2$ refers to collapse state (transaction system completely collapses and customers cannot finish transaction).

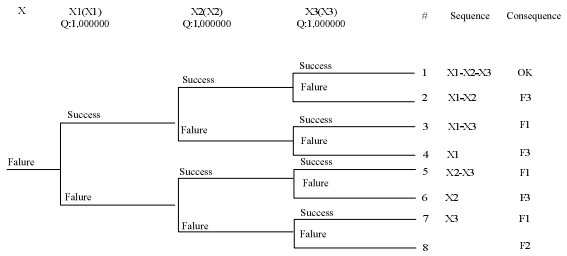


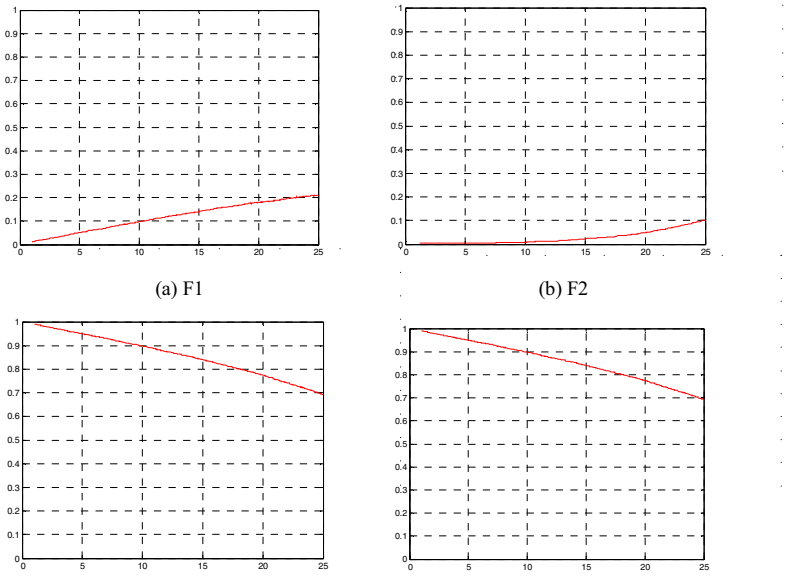
Figure 5. Event tree of the banking transaction system.

B. *Basic Safety Assessment*

According to the algorithms in above sector, the discrete-time Bayesian network is established. We set the task time $T = 1, 2, ..., 25$ and $n = T$, the timeline is separated as

$$T_n = \{[0,1], ..., [i-1, i], ..., [n-1, n), [n, \infty)\}$$

Based on the formulas in sector II, the probability of each consequence in different task time can be calculated. The key point lies in meticulous degree of timeline division. The more meticulous time-line divides, the higher calculation precision is, and the more resources calculation needs. When the memory of PC is 512MB and $n > 25$, it is prompted the memory is insufficient and the calculation can not be finished. When we adopt the improved method, even if $T > 100$, the task is still finished easily. Figure 6 shows the consequence probability of each results of P, changing with the time.



(a) F1 (b) F2

(c) F3 (d) OK

Figure 6. Consequence probability.

V. CONCLUSION

Contraposing to the disadvantages of traditional Markov chain analysis methods on dynamic fault tree, this paper studies transformation methods from "priority or gate", sequence correlation gate, functional correlation gate, spare parts gate of public spare parts and correlation gate of layer functions. Besides, we also studies the calculation methods of top event probability and the importance based on dynamic Bayseian network. Our scheme adopts the idea of integrated modeling and it analyzes transformation from fault tree and event tree to Bayesian network. It cannot only get general safety analysis results but it can also get other useful information, which is very convenient to make inference and diagnosis. Besides, above models and methods are applied to analyze the safety assessment of bank transaction system.

REFERENCES

[1] Boudali H., Dugan J.B., A discrete-time Bayesian network reliability modeling and analysis framework, Reliability Engineering and System Safety, vol.87, no.3, pp.337-349, 2005.

[2] Ren Jia, Tang Tao, Wang Na, Flexibility Discrete Dynamic Bayesian Networks modeling and Inference algorithm , Proceedings of the 24th Chinese Control and Decision Conference, pp.1675-1680, 2012.

[3] Xiao Zou, Heng Wang, Qiuyu Zhang, Hand Gesture Target Model Updating and Result Forecasting Algorithm based on Mean Shift, Journal of Multimedia, vol.8 ,no.1, pp.1-7, 2013.

[4] Tchangani Ayeley P., Noyes Daniel, Modeling dynamic reliability using dynamic Bayesian networks, Journal Europeen des Systemes Automatises, vol.40, no.8, pp.911-935, 2006.

[5] Zhou Zhongbao, Zhou Jinglun, Sun Quan, Dynamic fault tree analysis method based on discrete-time Bayesian networks, Hsi-An Chiao Tung Ta Hsueh/Journal of Xi'an Jiaotong University, vol.41, no.6, pp.732-736, 2007.

[6] Chung-Hung Tsai, The E-Commerce Model of Health Websites: An Integration of Web Quality, Perceived Interactivity, and Web Outcomes, Journal of Networks, vol.6, no.7, pp.1017-1024, 2011.

[7] Dugan J B, Bavuso S J, Boyd M A, Dynamic Fault-Tree Models for Fault-Tolerant Computer Systems, IEEE Transactions on Reliability,vol.41, no.3, pp.363-377, 1992.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

New Digital Thermostat Development

Wang Dong
Emerson Xi'an Engineering Center,
Xi'an, China
Davidw.Wang@emerson.com

Dai Xunjiang
Emerson Xi'an Engineering Center,
Xi'an, China
Xunjiang.Dai@emerson.com

***Abstract*—a new digital thermostat with features as Microcontroller unit-based, programmable, friendly-user interface, ultra-low power consumption, and "Green" has been developed. The design of hardware and software are discussed in detail in this paper. The hardware includes power supply circuit, MCU-based control circuit, voltage divider circuit and LCD display circuit; Software design is innovative as well, which includes temperature and humidity measurement and calibration algorithm, thermistor compensation algorithm, PWM signal modulation algorithm and PID algorithm so on. The purpose of digital thermostat research is to reduce power consumption and accomplish automatic fail-safe protection. In addition, cost analysis, bill of material, compliance tests are simply introduced to qualify the product design. Because of its pb-free electronic components and mechanical parts, plus power saving, the digital thermostat is a "Green" product, its performance and reliability are guaranteed by comprehensive hardware and software consideration. Therefore, the digital thermostat development will benefit to sustainable and green product development.**

Keywords-digital thermostat; hardware; software; green

I. INTRODUCTION

Based on the concept of new product [1], a new digital thermostat was developed which featuring a MCU as a key component, interface circuits and software algorithm. The digital thermostat can measure and regulate temperature accurately and precisely, to increase comfort and environmental quality for people's life. The temperature can be controlled digitally so the air conditioning system and heating elements will operate at the optimal level which greatly reduces the energy consumption. Therefore, the digital thermostat is a kind of "Green product" or "power saving product" in some sense.

II. HARDWARE CIRCUIT DESIGN

The hardware design is mainly focused on power supply circuit, temperature and humidity measurement circuit.

A. Power Supply Circuit

The power supply circuit is composed of current transformer, TRIAC, rectifier, SCR and electrolytic capacitor, which characterized as power saving mode. Its operating principle is described as follow: the input AC voltage is 120VAC or 240VAC, 50 or 60 Hz and the load power is assumed kilo-watt level. the TRIAC is triggered bi-directionally on and off by triggering signal, to supply current discontinuously for the resistive load; meanwhile the current transformer is used to power the low voltage circuit, the triggering technique is innovative: when the SCR is triggered on by trigger level from microcontroller, the rectified side of first bridge rectifier would be shorted, then TRIAC is triggered on automatically, so the power stealing mode through coordination of current transformer and TRIAC is accomplished. When the triggering pulse is PWM signal[2], the AC current flowing through the resistive load is indirectly modulated, so the load operation can be dynamically adjusted, the room temperature fluctuation will be controlled within ±0.1ºC by comparing ambient temperature and setpoint temperature to execute PID control algorithm. The schematic of power supply circuit is shown in Figure.1. The input is universal power supply 85Vac-265Vac, 50Hz and 60 Hz, its SELV(Secondary extra low voltage) is 3.3VDC system.

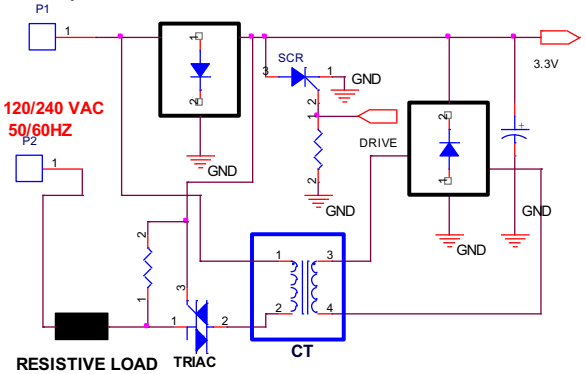


Figure 1. The schematic of power supply circuit

B. Temperature Measurement Circuit

There are many temperature measurement circuits that can be chosen, the practical circuits and interface technique will be thermocouples, resistive temperature detectors, thermistor and silicon integrated circuits and so on. The basic requirements of temperature sensor and conditioning circuit are to maximize the measurement accuracy and simplify the interface to the microcontroller. The available sensor interface options that are proportional to temperature include analog, frequency, ramp rate (dual slope ADC[3]), duty cycle, serial output and logic output. Due to the residential application of digital thermostat, thermistor circuit and analog output are considered for the design.

Voltage divider circuit is used to measure ambient temperature. The circuit consists of a voltage divider and a

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.101

voltage follower(buffer) op amp with a gain of one. Thermistors offer the advantages of a high sensitivity (R vs. temperature) and a linear change in resistance between approximately 0°C and 70°C. The voltage divider network consists of reference voltage VREF and series resistor RS. A low-pass, noise-reduction filter is formed by R1 and C1. The equation listed below can be used to select Rpull. Figure.2. shows the conventional circuit used with thermistors.

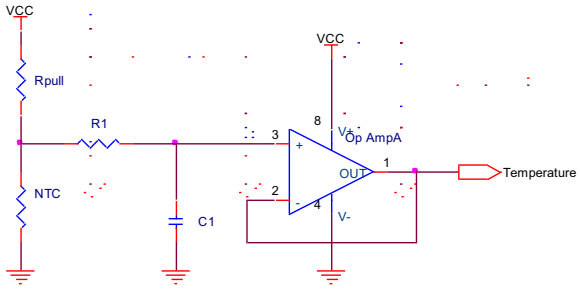


Figure 2. voltage divider circuit

$$R_{pull} = \frac{R_{TL}R_{Tm} + R_{Tm}R_{TH} - 2R_{TL}R_{TH}}{R_{TL} + R_{TH} - 2R_{Tm}} \quad (1)$$

Where: RTL= thermistor resistance at the low temperature.

RTm= thermistor resistance at the mid-point temperature.

RTH= thermistor resistance at the high temperature.

Regarding the voltage divider temperature sensing circuit, noise reduction technique should be paid attention for accurate temperature measurement. Therefore, proper grounding, EMI/ESD filter are required to prevent noise from degrading the accuracy of measurement. The noise can be induced into the measurement and the magnitude of the sensor voltage will be affected by ground bounce or switching noise at the amplifier ground. Because EMI or ESD overvoltage will likely occur, Ferrite beads, capacitive feed-through filters and RC filters, TVS etc should be taken into consideration. The RC filter only limit the slew rate of a transient-input voltage, A voltage-clamping device (such as a TVS zener diode) is required to limit the input voltage to a safe value that will not damage the IC amplifier.

C. *Humidity measurement circuit*

Humidity sensor SHT71x[4] is to be integrated into thermostat to measure the ambient relative humidity. The SHT71x is a single chip relative humidity module comprising a calibrated digital output. The device includes a capacitive polymer sensing element for relative humidity and a bandgap temperature sensor. Each SHT71x is individually calibrated in a precision humidity chamber. The calibration coefficients are programmed into the OTP(one time programming) memory. These coefficients are used internally during measurements to calibrate the signals from the sensors.

The 2-wire serial interface and internal voltage regulation allows easy and fast system integration. Its tiny size and low power consumption makes it the ultimate choice for even the most demanding applications.

The hardware block diagram of sensor and MCU is indicated in the Figure.3.

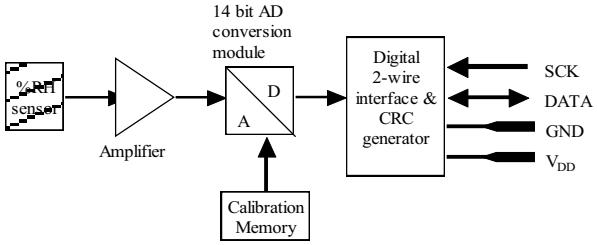


Figure 3. Humidity sensor hardware interface diagram

III. SOFTWARE ALGORITHM

A. *Temperature measurement algorithm*

Based on the voltage divider temperature measurement circuit, the R-T is non-linear, but the four pieces of linear curve can be used to fit the R-T curve, the Fig.4 is shown the R-T curve fitted by four linear curves. These four linear pieces intersect at three points along the curve. The slopes m of these four lines are pre-determined, as shown in the figure. To calibrate the thermostat, the location of the three points of intersection shall be defined at final test time and stored in non-volatile memory. In this case, thermostat calibration can be carried out to virtually any degree of precision, from a simple shift to a complete curve change.

The temperature can now be calculated using T=m·R+b, where T is the temperature (in °F), m is the appropriate slope according to what range R falls into and b is the appropriate y-intercept according to what range R falls into.

The measurements taken using thermistor values of RTH =75k, 150k, and 249k, will be referred to as R1, R2, and R3 respectively. This calibration scheme is designed to compensate for two sources of error: (1).The thermistors used are within ± 5% tolerance. (2).Other circuit variations, including operational amplifier differences, resistor tolerances, and mostly capacitor tolerance. These variations are compensated by testing the oscillator frequency at three fixed thermistor values as mentioned above, and fitting the curve appropriately.

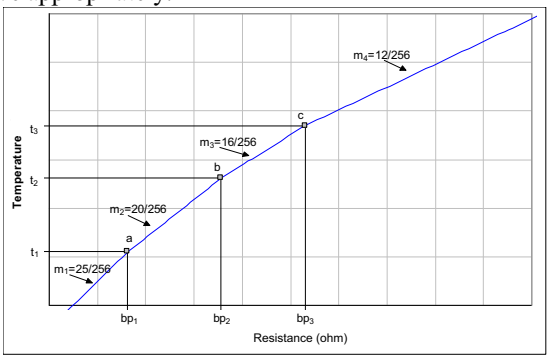


Figure 4. R-T fitted curve

Through lab testing of the thermistor, it has been determined that a good model for the temperature Vs resistance curve is given by:

$$R(A) = (1 + tol)(a + bA)^c \quad (2)$$

where tol = the thermistor tolerance
A = the ambient temperature, in °F
a, b, c=constant which obtained through lab testing
The inverse of this function or equation is:

$$A(R) = \left[\left(\frac{R}{1 + tol} \right)^{1/c} - a \right] \Big/ b \quad (3)$$

It is required is to find out how much a value variation in R affects the corresponding "temperature". Rather than solving this for all cases, we shall proceed by solving this at three points of interest; three calibration points. It is solved for -4.5%, nominal, and +4.5% with 75k, 150k, and 249k which gives the following values, using ΔR= -4.5%, 0, and +4.5%.

TABLE I. CALCULATED TEMPERATURE WITH THREE RESISTORS

R(A)	A,°F		
RTH, kO	*RTH-4.5%*	*Nominal*	*RTH+4.5%*
75	87.02	88.96	90.83
150	59.04	60.83	62.54
249	40.01	41.69	43.30

So this data can now be used to compensate for thermistor tolerance in the calibration process.

B. Humidity compensation algorithm

To compensate for the non-linearity of the humidity sensor and to obtain the full accuracy, it is recommended to convert the readout with the following two-order formula:

$$RH_{Linear} = C_1 + C_2 \cdot SO_{RH} + C_3 \cdot SO_{RH}^{\ 2} \quad (4)$$

Where SO_{RH}^2= sensor readout

TABLE II. HUMIDITY CONVERSION COEFFICIENTS

SO_{RH}	C_1	C_2	C_3
12 bit	−4	0.0405	-2.8*10-6
8 bit	−4	0.648	−7.2*10-4

The relation curve of sensor readout and relative humidity is shown in Figure.5

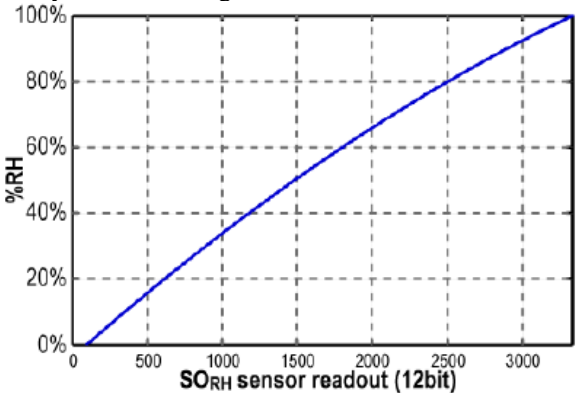


Figure 5. Sensor readout Vs relative humidity

As for Humidity Sensor RH compensation, For temperatures significantly different from 25 °C (~77 °F) the temperature coefficient of the RH sensor should be considered:

$$RH_{true} = (T_{\circ C} - 25) \cdot (t_1 + t_2 \cdot SO_{RH}) + RH_{Linear} \quad (4)$$

TABLE III. CONVERSION COEFFICIENT CONSIDERING THE TEMPERATURE EFFECT

SO_{RH}	t_1	t_2
12 bit	0.01	0.00008
8 bit	0.01	0.00128

TableIII is about the conversion coefficients of humidity compensation algorithm considering the temperature effect.

C. PID Controller Algorithm

PID control algorithm is widely used in temperature control system. Figure 6 shows a basic block diagram of a feedback PID control system.

The proportional term is the simplest of the three and is also the most commonly found control technique in a feedback system. The proportional gain (K_p) is multiplied by the error. The amount of correction applied to the system is directly proportional to the error. As the gain increases, the applied correction to the Plant becomes more aggressive. This type of Controller is common for driving the error to a small, but non-zero value, leaving a steady state error.

Integral control (K_i) looks at past errors, the accumulative error (sum of all past errors) is used to calculate the integral term, but at fixed time intervals. A temperature system would require a longer sample period than a motor system because of the sluggish response in a temperature controlled environment. If the integral sample period was too fast in the temperature system, the accumulative error would add too quickly to give the system a chance to respond, thereby not allowing it to ever stabilize. Therefore, as the accumulative error increases, the integral term has a greater effect on the Plant. In a sluggish system, this could dominate the value that is sent to the Plant.

Derivative term makes an adjustment based on the rate at which the Plant output is changing from its Setpoint. A notable characteristic in this type of control is when the error is constant, or at the maximum limit, the effect is minimal.

Finally, PID tuning is required to find the optimized PID constant, namely K_p, K_i, K_d. however, Tuning a PID Controller can be somewhat difficult and time consuming and should be completed in a systematic fashion.

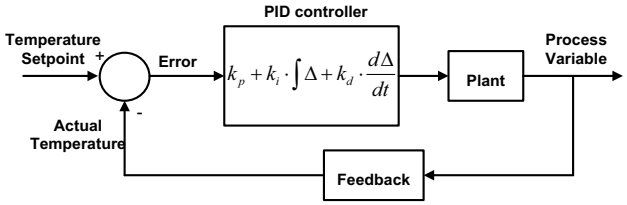


Figure 6. Basic block diagram of a feedback PID control system

IV. SIMULATION AND EXPERIMENTAL TEST

A. Simulation Analysis

ORCAD and Pspice[5] are used to do simulation analysis on the analog & digit mixed circuit. The critical of simulation on the power supply circuit is to build an accurate Pspice model. The Figure.7 is how that the load current is dynamically modulated by the triggering PWM signal. Figure.8 is the secondary side DC voltage output.

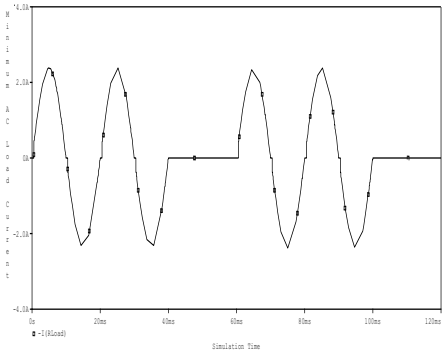


Figure 7. Simulation result for minimum AC load current

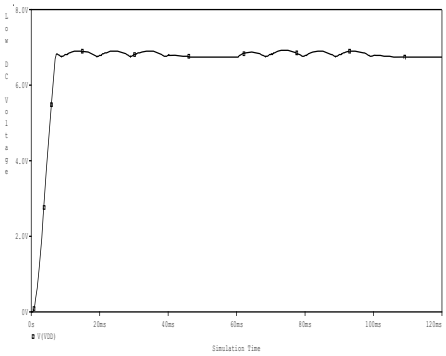


Figure 8. Simulation result for low DC voltage

B. Experiment test

Target board was built to test hardware and debug software, below figures show waveforms probed through oscillator, which further verify the correctness and feasibility of hardware. Fig.9~Fig.11 showed captured waveforms.

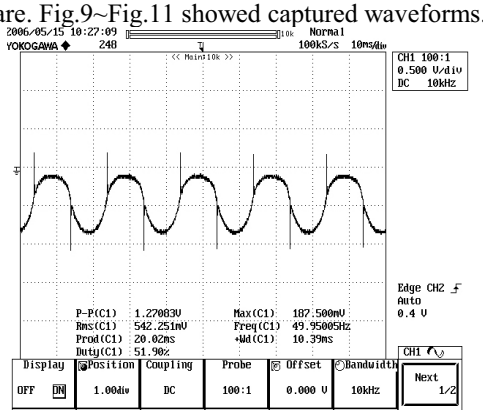


Figure 9. The waveform between gate and MT1

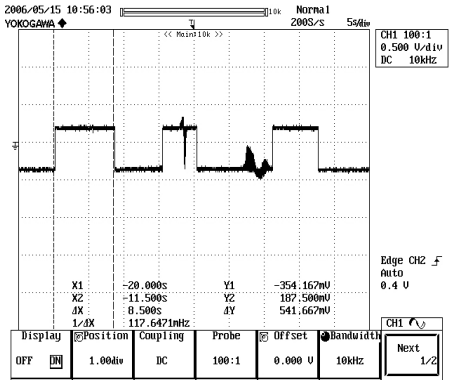


Figure 10. PWM triggering signal

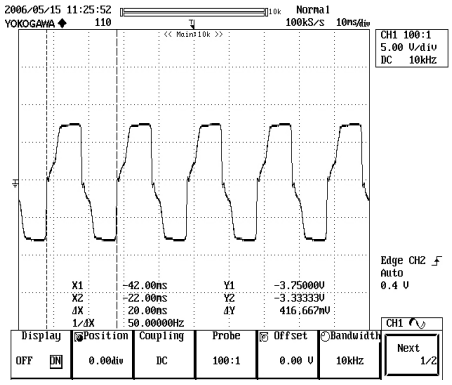


Figure 11. Current transformer secondary side output

V. CONCLUSION

The design of digital electronic product development has the super advantages of short development period, advanced technology and fast shift to market change. Hardware and software design of thermostat are given in detail, finally the simulation results and experimental test waveforms are compared as well. The target is to seek the minimum of power consumption and cost, maximum of power saving, meanwhile to care more about the users for convenience and safety, this is an inevitable trend of technology development in the age of competitive electronic product. The ultimate goal is to accomplish product design as possible and launch it into market worldwide successfully, in addition to providing "Green" electronic product for people.

REFERENCES

[1] RockyR.Arnold,Ph.D. New Product Development Challenge of Design for Environmental Compliance and Design for Electromagnetic Compliance. IEEE, pp.287-291,2004

[2] Guan-Chyun Hsieh Jin-Fu Yan. Group-Asymmetrical PWM Controller for Dimmable Fluorescent Lamp Ballast without Striation and Thermostat Effect. IEEE,pp.792-797,2005

[3] TI datasheet of microcontroller chip MSP430.

[4] Datasheet of Humidity sensor of SHT71x.

[5] Orcad and pspice user guide.

Design of Epidemic monitoring platform based on ArcGIS

Zhang Mei
Information Institute
GUI Zhou University of Finance and Economics
Guiyang, 550004, China
E-mail: zm_gy@sina.com

Yang Zirong
Information Institute
GUI Zhou University of Finance and Economics
Guiyang, 550004, China
E-mail: 277879478@qq.com

***Abstract*—With rapid development of IT, GIS technology in the field of preventive medicine has been an unprecedented application. Emerging epidemics of the 21st century pose a serious threat to security of human life, strengthening epidemiological surveillance, early warning and forecasting and to take timely preventive and control measures rapidly have very great significance. In this paper, combine GIS spatial analysis with epidemiology, choice "SARS" as the research object, it is designed that the development framework, main function, interface, spatial database and property Database of the surveillance platform.**

Keywords-ArcGIS; epidemics; surveillance; SARS; function structure design; database design

I. INTRODUCTION

At present, China's public health information system lacks a unified platform, poor timeliness of disease surveillance, data integration is weak, long epidemic reporting period, in the health information network covering a small surface, the lack of case reports, information collection is not comprehensive enough, so that our country in the face of major emergencies very passive when public health events [1]. GIS technology developed in recent years to study the spatial distribution of the epidemic and its related issues offer new solutions and tools [2]. Especially GIS based spatial analysis technology to expand the space-time decision-making platform, providing more advanced models and tools to solve these problems [3].

World Health Organization will organize the internal management system, through this system will store all information contact spatial information of various types of infectious diseases in the world together, when an outbreak of disease can analyze the epidemiological characteristics of confirmed cases and the implementation of location tracking, timely prevention and treatment [4]. In 2003, various countries of the world have set up corresponding control epidemic early warning monitoring system after the global outbreak of severe acute respiratory syndrome (SARS) epidemic. In the UK, the government established a national health information and monitoring service-based system using GIS technology [5]. In the United States, they established a public health surveillance system, but the system can do real-time monitoring of infectious diseases, the development of automatic monitoring of the epidemic, and promptly report the epidemic of deaths data. In our country, the same year launched a national disaster prevention and emergency public health time reporting system [6], the timely collection of all kinds of unexpected domestic public health time, a comprehensive summary of the event information, and the results will be reported promptly to the state and all levels of government related to the health sector, the Centers for Disease control and hospitals make prevention and treatment programs as early as possible, as soon as possible to control the epidemic. However, there are some flaws in this system, not in-depth and comprehensive information on mining, it can only report outbreaks event, for certain diseases cannot report a case by case basis, but also a lack of decision analysis capability after information conformity, it is difficult for decision-makers of health prevention and control station provide information of decision support.

Lu et.al [7] discussed GIS-based system of SARS emergency medical system design and architecture. Chinese Academy of Sciences developed a SARS control and early warning information system [5], published on the website of the Ministry of Health, while its SARS epidemic in Beijing has established control early warning information system. PLA Information Engineering University to establish a SARS epidemic in Henan Warning Decision Support System [8].

II. THE OVERALL DESIGN

A. Development Framework

SARS epidemic monitoring platform is a full use of technology, geographic information systems, database technologies and the SARS epidemic epidemiology and computer data processing platform constructed space. The platform architecture based on B / S carried out large-scale industry database server using SQL Server2005 add spatial data engine ArcSDE of Geodatabase, which SQL Server2005 database management attributes, Geodatabase, by manipulating spatial data spatial data and table data conversion and storage of different types of databases connection data, enables the exchange of data with the foreground application [9] .It shows the system B \ S architecture diagram in Figure 1.

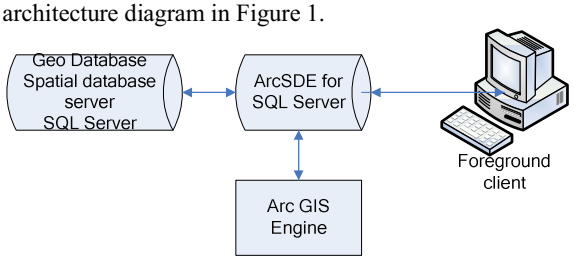


Figure 1. System B \ S architecture diagram

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.102

System database consisting of the spatial and attribute database [10]. Among them, the system database storage space design all kinds of vector data, raster data, such as map data. By spatial data engine ArcSDE Geodatabase - meaning all types of spatial data and be expressed. Property database is used to store the epidemic process information related to suspected cases, confirmed cases of information, hospital information, PFA information, quarantine information, attribute data; this property category to establish a database to store various types of data, to protect the stability of database systems, security ability to access and efficiency.

B. *Main Function of Design*

As showed in Figure.2, this platform has the following features:

(1) SARS information collection: gathering information SARS cases, SARS cases integrated into the repository, build property database of information under the SARS epidemic.

(2) SARS epidemic and related information management: pre-SARS case information collected, SARS cases in close contact info, quarantine information, PFA information, SARS designated hospitals and health point of resource information for routine management add, modify and query .

(3) SARS outbreak analysis. According to the spread of SARS epidemic characteristics combined with GIS, we see the distribution of the SARS epidemic and the like on the map, and in accordance with the regional distribution or \ day the number of deaths, make the appropriate statistical analysis table;

(4) Analysis of disease transmission mode. According to patient information, combined with daily location surrounding circumstances surrounding the case, establish SARS cases roadmap activities, propagation path of patient contacts, and displayed on the map.

(5) The results of the analysis to guide the implementation of space-related measures: According to the current location of the patient before the onset of the active area and use GIS to share space on the principle of combining health resources at its disposal situation, establish SARS cases and activity diagrams may infected area, to determine the best solution for public health resources, including the shortest route, the most appropriate hospitals to accept patients, and health and epidemic prevention station and so on.

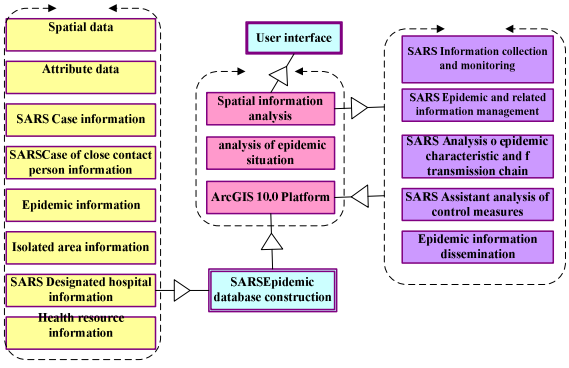


Figure 2. Functional structure design

C. *Interface Design*

Interface is a platform by which a client can achieve visually information interaction and operation through computer. It is an important part of software design. Outstanding software products generally have a good interface. Interface design always follows the following principles: Firstly, make sure good interaction between users and software in a controlled environment ; Secondly, the interface style should be consistent overall program including software products' framework , object-oriented design principles and safety requirement of design product ; Thirdly, all levels interface should be kept as simple as possible , beautiful and good visual effective. The SARS epidemic monitoring platform follows these principles to design a user-friendly, beautiful, features-clearly, practical and convenient main interface, shown in Figure 3.

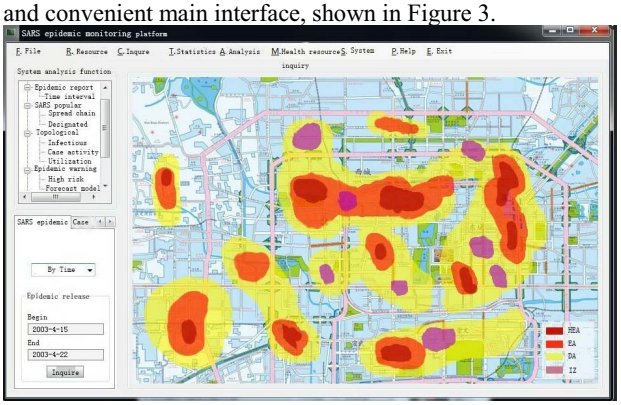


Figure 3. Main interface platform

III. DATABASE CONSTRUCTION

A. *Space Database Structure*

Spatial database is a set of description, storage and processing spatial data, which is narrowly defined as map data management. It is one of the important part of GIS, playing a vital role in the geographic information system. It combines spatial data collection, storage, retrieval, and analysis together, can be retrieved according to the needs of different departments and provide the reliable and accurate basis for the manager's decision making. Classified by the data types stored in the spatial database, it can be divided into five categories as follow: vector data, digital elevation model, digital orthogonal projection image, digital raster map and thematic data, as shown in Figure 4.

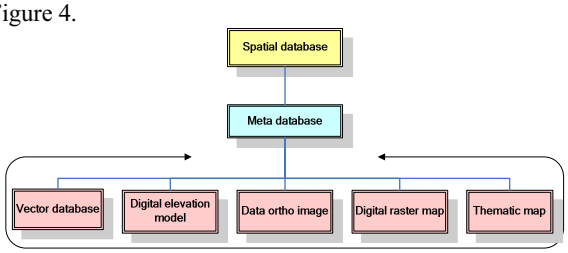


Figure 4. Spatial database structure

B. *ArcSDE Geodatabase*

ArcSDE is often used as the interface to connect to geographical database. Any spatial data can be fixed by ArcSDE to load into the relational database, it can be said ArcSDE and ArcGIS as a bridge between relational databases. Moreover, many other data management system also can use it to realize the management, while making the ArcGIS call any data [11].ArcSDE has two components, namely the spatial analysis tools and geographic data service. Spatial data united ArcSDE and RDBMS together by geographic data, assisting it complete data storage and release. All kinds of database tables can be stored in RDBMS in the form of physical and ArcSDE is responsible for the timely information of all the tables for the GIS.

As mentioned above, ArcSDE is a set of spatial database management software designed by ESRI Company for spatial data storage issues. In order to obtain efficient spatial data management, and stored in the database, users can unify the different data product supported by ArcSDE into Geodata model.

The relationship between ArcSDE and Geodata is shown in figure 5. At present, increase the spatial data engine based on the traditional relational database is relatively popular in the spatial data management solution. In order to store and manage direct vector spatial data directly, some database management systems software vendors extended their own systems. But the principle of the two expansion mode is the same, both using coordinate data in the database management system BLOB field storage space object.

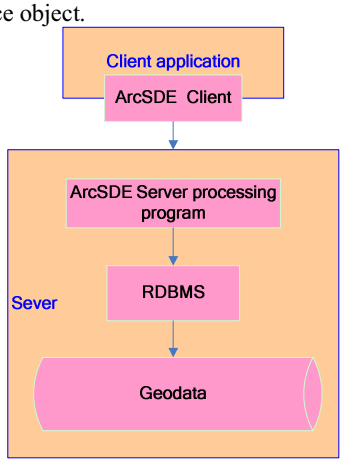


Figure 5. Relationship between ArcSDE and Geodatabase

C. *Data Sources*

Relying on this platform application project in Guizhou Province Natural Technology Fund project, which involves a large number of SARS infected areas and epidemic information, quarantine information, information on case, close contact info, SARS treatment hospitals and health stations and other data, these data There are third project application Yunnan Center for Disease Control and Prevention, Professor of Humanities Weihua Wen where available. Yunnan Provincial Disease Prevention and Control Center was established in December 18, 2001, is a public good implementation of disease prevention and control and public health technology management and services organized by the Yunnan Provincial People's Government and institutions, the Centre has initially built a complete professional, advanced equipment, Sophisticated technology, strong comprehensive ability of the province's health and disease control CDC technical guidance center. Currently, the center of AIDS, cholera, SARS (SARS), influenza A (H1N1), the disposal of public health emergencies, health food and cosmetics, health inspection, a number of professional management and disinfecting prevention and control of vector or areas in the national advanced level. He won the Central Organization Department awarded the "National Working prevention of atypical pneumonia advanced grassroots party organizations," the title. Professor Wen-Wei Hua has been working at the center, the prevention and control of SARS epidemic have experienced since July 2003, be able to effectively use its Epidemiology experimental conditions based platform for all data acquisition and processing, disease research provided stand by.

D. *Property Database Table Structure*

Use SQL Server 2005 software to establish five SARS outbreak-related information and data classification. In order to enable it to be fast, and so on classification, query and related performance data in accordance with the five categories of information on demographic characteristics, they are divided into 15 data tables. Meaning database tables as shown in Table 1.

TABLE I. MEANING OF DATABASE TABLES

Table Name	Table name Meaning
MILITARY_YBQK	SARS cases in general
MILITARY_FBJZ	SARS cases and Consultation
MILITARY_LCJC	Clinical inspection of SARS cases
MILITARY_LXBX	investigation of SARS cases
MILITARY_ZDXJ	SARS Case Study Summary
M_CONTACY_YBQK	close SARS cases in general
M_CONTACY_JCDD	Close contact with SARS exposure sites
M_CONTACY_JCFS	Contact manner with SARS
BYXJG	Pathogen Detection
JCDW	Contact with animals case
JCFD	Contact with SARS patient's condition
JCRY	The contact person
GLQ_INFO	Quarantine Information
ZY_INFO	Health Resources Information
YQ_INFO	PFA Information

Space restrictions, the following are a few typical database table structure, as shown in Table 2 is SARS and the contact person table JCRY. Table 3 is SARS close contacts general table M_CONTACT_YBQK.

TABLE II. MEANING OF DATABASE TABLES

Field name	Data Types	Explanation
JCRY_ID	Digital	Case Coding
JCRY_RQ	Date	Date
JCRY_HDNR	Text	Activities
JCRY_HDDD	Text	Activities Location
JCRY_FR	Text	Fever human contact
JCRY_NFR	Text	No heat human contact

TABLE III. SARS IN CLOSE CONTACT WITH GENERAL FACT SHEET M_CONTACT_YBQK

Field name	Data Types	Explanation
C_XJZD_XIAN	Text	County
C_XJZD_XIANG	Text	Township
C_PHONE	Text	Phone
C_GZDW	Text	Workplace
C_GLFS	Digital	1 = worse 2 = medical 3 = kept for station; 4 = no quarantine
C_GLSJ	Text	Isolation time
C_ZG	Digital	1 = desegregation; 2 = Switch suspected; 3 = Switch clinical diagnosis; 4 = out of isolation; 5 = death; 6 = Other
C_ZG_RQ	Text	Vesting date
C_ZG_SFZL	Digital	1 = Yes; 2 = No
C_ZG_SFZL_YY	Text	Name of hospital treatment
C_ZG_ZLBQ	Text	Isolation and treatment start time
C_ZHJCRQ	Text	The last contact time
C_DCDW	Text	Investigation Unit
C_DCRQ	Text	Survey Date
C_DCZQM	Text	Investigators signature
C_SHEARCH	Logic	1 = visited ; 2 = unvisited

IV. CONCLUSION

Build Beijing-based SARS outbreak data monitoring platform to meet the basic national health administrative departments of the public emergency response and control needs. The platform of the SARS epidemic information collection and monitoring, building the SARS case information database table, you can always check the SARS epidemic changes and scientific analysis and assessment of the extent and scope of the spread of the epidemic, SARS epidemic can be within the specified geographical area Popular features for qualitative analysis, mainly in accordance with the infected person's age, gender, occupation and other classification, location and area classifications according to the affected areas, according to the spread of the epidemic development time feature classification, and draw statistics. The platform can be set up in accordance with the chain of transmission of SARS epidemic and in close contact with the activities, while the reverse analysis, traced the source of infection transmission chain formed by existing SARS cases or close contacts. The platform can be based on the patient location and the onset of the first two weeks, a week to the hospital and diagnosed with the venue and the scope of the three time periods, the establishment of case activity roadmap delineate the area likely to be infected, and respond to warning signs, so as soon as possible the preventive measures. The platform can also be based on the patient's location, combined with the distribution of the peripheral hospitals or health assistance possible point and can be utilized to determine the best treatment program.

ACKNOWLEDGMENT

Thank Project Supported by Regional Science Fund of National Natural Science Foundation of China (Project approval number: 41261094. Thank Project Supported by Nomarch Fund of Guizhou Provincial excellence science and technology education person with ability (No: Qian Province ZhuanHeZi (2012)156). Thank Project Supported by Natural Science Research Yonth foundation of Guizhou Provincial Department of Education (QianZhuanHeKzZi (2012)074).

REFERENCES

[1] Liu-Jiyuan,Zhong-Ershun,Zhuang-Dafang,Wang-Jingfeng,Song-Guanfu . Development and Application of National SARS Disease Controlling and Pre-warning Information System[J] . Journal of Remote Sensing,2009.9(7):337-344

[2] Wang-Zheng,Li-Huaqun,Chen-Jianguo,Cai-Di,Li-Shan,Wang-Ying,Zheng-Yiping,Wu-Bing . Emergency management system for SARS prevention[J] . Journal of Remote Sensing, 2005.6(5):82-85

[3] Tang-Guoan,Yang-Xin . Experimental Course of ArcGIS Geographic Information Systems Spatial Analysis [M]. Science Press,2012

[4] Meng-Lingkui,Shi-Wenzhong,Zhang-Penglin,Huang-Changqing . Network GIS theory and technology[M]. Science Press,2010

[5] Liu-Jiyuan,Zhong-Ershun,Zhuang-Dafang,Wang-Jingfeng,Song-Guanfu . Development and Application of National SARS Disease Controlling and Pre-warning Information System[J] . Journal of Remote Sensing, 2003,7(5):337-340

[6] Dong-Zeyu . Development Process of American Warning Systemand Its Enlightenment[J]. China Public Security(Academy Edition),2014,2:1-5

[7] Lu XL.A WEB-GIS Based Urgent Medical Rescue CSCW System for SARS Disease Prevention[J].LNCS,2004;3032:91-98

[8] Nayan Sharma;R. D. Garg;Archana Sarkar . RS-GIS Based Assessment of River Dynamics of Brahmaputra River in India [J] . 2012, 4(2):113-119

[9] Tyler J. Tran and Katherine J. Elliott.Estimating Rhododendron maximum L. (Ericaceae) Canopy Cover Using GPS/GIS Technology[J].2012,77(4):245-251.

[10] Wu-Jing,He-Bi,Li-Haitao . Geographic information system application tutorial to ArcGIS 9.3 desktop [M] . Tsinghua university press,2011

[11] Maged N Kamel Boulos . Towards evidence-based,GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom[J] . Int J Health Geogr, 2004,3(l):l-30

Improvement of Dynamic Time Warping (DTW) Algorithm

Yuansheng Lou
College of Computer and Information
Ho Hai University
Nanjing,China
wise.lou@163.com

Huanhuan Ao*, Yuchao Dong
College of Computer and Information
Ho Hai University
Nanjing,China
*Corresponding author: aohuanhuanhhu@163.com

Abstract: **Dynamic time warping algorithm is widely used in similar search of time series. However, large scales of route search in existing algorithms resulting in low operational efficiency. In order to improve the efficiency of dynamic time warping algorithm, this paper put forward an improved algorithm, which plans out a three rectangular area in the search area of the existing algorithms, search path won't arrive the points outside the rectangular area, thus further reduced the search range of regular path. To some extent, this algorithm reduces computation of original algorithm, thus improving the operational efficiency. The improvement is more pronounced when two time series are longer.**

Keywords: dynamic time warping; time series; similar search; warping path; Euclidean distance

I. INTRODUCTION

In the early 1970s, Japanese scholars Itakura put forward dynamic time warping algorithm (DTW), introducing the concept of dynamic programming into the hard problem of recognition of the talk with isolated words in an uneven speed, thus significantly improving the efficiency of the recognition of isolated words [1]. With the spring up of time series data mining, the concept of DTW has been introduced in similar search field of time series by scholars at home and abroad [2]. After large amount of experiments, great achievements are made. The idea of DTW algorithm is a procedure using dynamic programming techniques to solve the optimization problem, during which a complex global optimization problem is broken into several local optimization problems, then decisions are made step by step [3], finally an optimal solution of the global problem is gotten. The procedure of time series similarity matching makes use of local optimization to find a path, along which he cumulative bending distance between two time series is minimum. But during the measurement of distance between two time series with DTW, it is vulnerable to be interfered by “noise" and "outlier" [4] of time series, since DTW algorithm matches point by point. And when the time series are too long, the calculation is very large. Therefore a suitable constraints and an appropriate matching range is the key to improve DTW matching accuracy and to shorten time of matching.

Based on analyzing existing DTW algorithm [5], this paper plans out three rectangles area outside the parallelogram path search range. The lattice outside the rectangular does not appear in the regular path; therefore there is no need to make a judgment that whether the lattice is within the area of parallelogram, which in turn decrease the calculation and increase the operation efficiency in some degree.

II. DTW ALGORITHM

A. Dynamic Time Warping Distance

Definition 1: Assume there are two certain length of time series R and T, of m and n. And$R = (r_1,\ r_2,\ r_3, \dots,\ r_m)$, $T = (t_1,\ t_2,\ t_3, \dots,\ t_n)$. In order to align these two sequences in time, firstly rectangle D with n row and m column need to be created [6]. Each matrix element $D_{ij} = d(r_i,\ t_j)$ represents the distance between the point r_i of series R and the point t_j of series T. This distance is usually Euclidean distance $d(r_i,\ t_j) = (r_i - t_j)^2; (1 \le i \le m, 1 \le j \le n)$. The shorter distance indicates the more similar between points, on the contrary, the longer distance indicates the less similarity [7].

$$D = \begin{bmatrix} d(r1, tn) & \cdots & d(rm, tn) \\ d(r1, tn-1) & \cdots & d(rm, tn-1) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ d(r1, t1) & \cdots & d(rm, t1) \end{bmatrix}. \quad (1)$$

In the distance matrix D, solve a successive collection of matrix element with dynamic programming algorithm. The connection of each element is called regular path W, $W = w_1, w_2, w_3, \dots, w_k, \dots, w_K$. The k-th element of W $w_k = (i,\ j)_k$ is the alignment of i-th point of series R and j-th point of series T. The distance of regular path W is the warping distance of dynamic time. Regular path is not chosen at random, it must meet the following constraints [8-9]:

a) ***Boundedness****: The length of regular path W should within this range* $max(m, n) \le K \le m + n - 1$*;*

b) ***Boundary conditions****: The starting point of regular path W is* $w_1 = (1,\ 1)$*, and the end point is* $w_K = (m,\ n)$*;*

c) ***Continuity:*** *Suppose the previous point* $w_{k-1} = (i', j')$*, next point* $w_k = (i,\ j)$ *in regular path, then there must be* $(i - i') \le 1$ *and* $(j - j') \le 1$*;*

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.103

d) ***Monotonic:*** *Suppose the previous point* $w_{k-1} = (i', j')$*, next point* $w_k = (i, j)$ *in regular path, then there must be* $0 \le (i - i')$ *and* $0 \le (j - j')$*;*

Combining the continuity and monotonic of these constra ints, if regular path W passes through the point (i, j), then the previous point must be one of point $(i - 1, j)$, $(i - 1, j - 1)$ and $(i, j - 1)$ .The number of regular path which meets the above constraints can be exponential, but the one we need is the one with least cost to regular following formula:

$$\mathrm{DTW}(\mathrm{R},\mathrm{T}) = \min\left\{\sqrt{\sum_{k=1}^{K} w_k}/K\right\} \tag{2}$$

B. Classic DTW Algorithm

In order to avoid unnecessary path search during DTW algorithm, the actual search path is limited at a parallelogram of 1/2 to 2 slopes in literature [10]. As shown below, m and n are two time series in parallelogram OABC. As $k_{OA} = 2$, $k_{OC} = 1/2$ are known, the functional relation of four sides and the coordinates of four points can be determined as follows: $y_{OA} = 2x$, $y_{OC} = 0.5x$, $y_{AB} = 0.5x + m - 0.5n$, $y_{CB} = 2x + m - 2n$, point $O(0,0)$, point $A(\frac{2m-n}{3}, \frac{2(2m-n)}{3})$, point $B(n, m)$and point $C\left(\frac{2(2n-m)}{3}, \frac{(2n-m)}{3}\right)$. x_a $(x_a = \frac{2m-n}{3})$ and x_c $(x_c = \frac{2(2n-m)}{3})$ are two nearest integer point, therefore the constraints for the length of m and n are $2m - n \ge 3$ and $2n - m \ge 2$. If these two constraints cannot be satisfied, then these two series cannot be dynamic warped due to the big difference.

The computation of distance matrix D is very large. When the search path is limited at parallelogram OABC, there is no need to calculate the matching distance of the lattice points outside OABC. Therefore the computation is cut down largely [11-12].

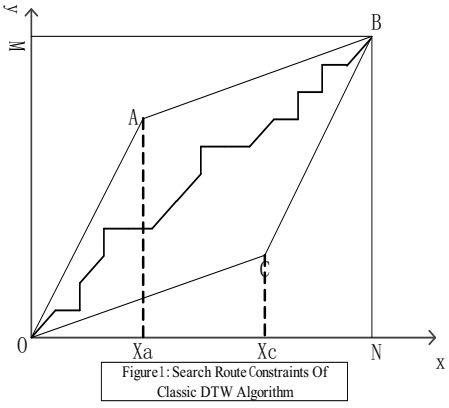


Figure 1: Search Route Constraints Of Classic DTW Algorithm

III. ALGORITHM IMPROVEMENT

A. Problems of Current Algorithms

In current DTW algorithm, distance matrix between two time series and accumulated distance matrix are need to be calculated, so the calculation is very large originally. Even if the search path is limited in a parallelogram, upper and lower boundary calculation is needed at each link of path search. The calculation is still very large, especially when two time series is very long, then this kind of repeated calculation is even more and the calculating efficiency is significantly reduced.

B. Principle of Algorithm Improvement

In order to avoid n*m times calculation of points' coordinates, firstly it needs to quickly determine which points are within OABC. Plan three rectangular areas Ω_1, Ω_2, and Ω_3 in figure 2. In the analysis, Ω_1 is determined by point $O(0,0)$ and point $A'\left(\left\lceil\frac{2m-n}{3}\right\rceil, \left\lfloor\frac{2(2m-n)}{3}\right\rfloor\right)$ which is the nearest integer point at top of A. Vertical line A'E intersects OC at point E' .Point $E\left(\left\lceil\frac{2m-n}{3}\right\rceil, \left\lfloor\frac{2m-n}{6}\right\rfloor\right)$ is the nearest integer point below point E'. Point $C'\left(\left\lceil\frac{2(2n-m)}{3}\right\rceil, \left\lfloor\frac{2n-m}{3}\right\rfloor\right)$ is the nearest integer at right of point C. The vertical line through point C' intersects AB at point D'. Point $D\left(\left\lceil\frac{2(2n-m)}{3}\right\rceil, \left\lceil\frac{4m+n}{6}\right\rceil\right)$ is the nearest integer point above point D'. Point D and point E can determine Ω_2. Point $C'(\left\lceil\frac{2(2n-m)}{3}\right\rceil, \left\lfloor\frac{2n-m}{3}\right\rfloor$ and point $B(n, m)$ can determine Ω_3. Firstly, the set of points in parallelogram OABC the set of points within $\Omega_1+\Omega_2+\Omega_3$ the set of points within rectangular OMBN, so the points outside $\Omega_1+\Omega_2+\Omega_3$ are not involved in the calculation of slope range. Most of points which are not on regular path are filtered. In other words, to locate points in OABC accurately means to find the set of points in $\Omega_1+\Omega_2+\Omega_3$ which meet the constraints as follows:

$$\begin{cases} y - 2x \le 0 \\ y - 0.5x \ge 0 \\ y - 0.5x - m + 0.5n \le 0 \\ y - 2x - m + 2n \ge 0 \end{cases} \tag{3}$$

Since not all the set of points in OMBN are involved in the calculation of slope range， the calculation is reduced largely that in turn improve the calculation efficiency. Especially when time series are longer, the improvement of calculating speed is even more evident.

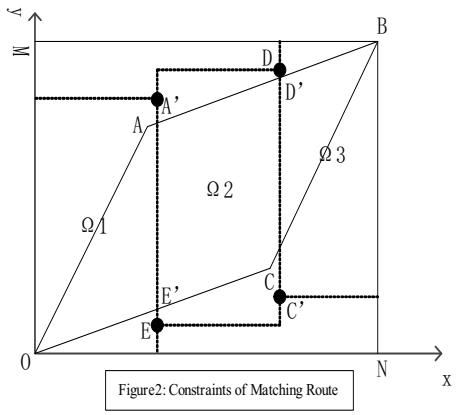


Figure2: Constraints of Matching Route

C. *The realization of improved algorithm*

Take time series $R = (r_1, r_2, r_3, \dots, r_m)$ and $T = (t_1, t_2, t_3, \dots, t_n)$ as input, improved DTW algorithm is realized as follows table I:

TABLE I. DTW(R, T) ALGORITHM

DTW(R, T) algorithm:

```
m = R.length; n = T.length
d, D = new[n × m]
warpingDistance = 0.0
for i = 0; i < ⌈(2m − n)/3⌉; i++
        for j = 0; j < ⌈2(2m − n)/3)⌉; j++
            if (i, j) in OABC
                d[i][j] = (R[i] – T[j])²
for i = ⌈(2m − n)/3⌉; i < ⌈2(2n − m)/3)⌉; i++
        for j = ⌊(2m − n)/6⌋; j < ⌈(4m − n)/6⌉; j++
            if (i, j) in OABC
                d[i][j] = (R[i] – T[j])²
for i = ⌈2(2n − m)/3)⌉; i < n; i++
        for j = ⌊(2n − m)/3⌋; j < m; j++
            if (i, j) in OABC
                d[i][j] = (R[i] – T[j])²
for i = 0; i < n; i++
        for j = 0; j < m; j++
            if d[i][j] == 0.0
                continue;
            if i == 0 && j == 0
                D[i][j] = d[i][j];
            elseif i == 0 && j != 0
                D[i][0] = d[i][0] + D[i - 1][0];
            elseif j == 0 && i != 0
                D[0][j] = d[0][j] + D[0][j - 1];
            else
                warpingDistance = min(D[i - 1][j] == 0.0 ? : +∞, D[i -
1][j - 1]) == 0.0 ? : +∞, D[i][j - 1]) == 0.0 ? : +∞)
                warpingDistance += d[i][j];
        D[i][j] = warpingDistance;
return warpingDistance
```

In the improved algorithm above, d[n][m], D[n][m] are respectively distance matrix and cumulative distance matrix. The lattice point outside the three rectangular in the figure 2 are not be calculated, and algorithm eventually outputs warping distance which is the distance of regular path, that is the shortest warping distance required by DTW algorithm.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental environment: Eclipse + java, win7;

System parameters: P320 AMD Athlon、2G memory;

Experimental data: In this experiment, the data is provided by Professor Eamonn Keogh of University of California. The website address: http: //www.cs.ucr.edu/ ~eamonn/time_series_data/. Select 15.dat data from the folder mariohWords as the experimental data.

Experimental procedure: There are 1861 pairs of data in sample data set. In order to compare efficiency of arithmetic of the conventional DTW algorithm with the improved algorithm with different lengths of time series, select samples as follows, sample 1: m = 100, n = 120; sample 2: m = 500, n = 540; sample 3: m = 1000, n = 1500; sample 4: m = 1500, n = 1800.Every two time series were tested five times and calculate average test results in both arithmetic. Results are shown in table II:

TABLE II. COMPARISON OF OPERATIONAL TIME IN TWO ALGORITHMS

Length of time series	Traditional DTW (ms)	Improved DTW (ms)	Improvement rate (%)
100×150	15.6	15.2	2.56
500×550	31.2	29.8	4.49
1000×1500	143.2	135.3	5.38
1500×1800	188.4	177.6	5.73

By table II: When the length of time series is 100 × 150, the improved algorithm saves 2.56% time than the traditional method. When the length is 500 × 550, the improved algorithm saves 4.49%. When the length is 1000 × 1500, the percentage is 5.38%. When the length is 1500 × 1800, the improvement rate is 5.73%. Results prove that improved algorithm proposed in this paper, to a certain extent, reduces the amount of computation and improve the operational efficiency and the improvement is even more obvious when the time series are long.

ACKNOWLEDGMENT

The research of this paper is partially supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under Grant No. 2013BAB06B04, No. 2013BAB05B00 and No. 2013BAB05B01.

REFERENCES

[1] Yuxin Z, Miyanaga Y. An improved dynamic time warping algorithm employing nonlinear median filtering[C]//Communications and Information Technologies (ISCIT), 2011 11th International Symposium on. IEEE, 2011: 439-442.

[2] Berndt D J, Clifford J. Using Dynamic Time Warping to Find Patterns in Time Series[C]//KDD workshop. 1994, 10(16): 359-370.

[3] Xu L, Ke M. Research on isolated word recognition with DTW-based[C]//Computer Science & Education (ICCSE), 2012 7th International Conference on. IEEE, 2012: 139-141.

[4] Wang Y, Lei P, Zhou H, et al. Using DTW to measure trajectory distance in grid space[C]//Information Science and Technology (ICIST), 2014 4th IEEE International Conference on. IEEE, 2014: 152-155.

[5] Islam S, Hannan M A, Basri H, et al. Solid waste bin detection and classification using Dynamic Time Warping and MLP classifier[J]. Waste Management, 2014, 34(2):págs. 281-290.

[6] Rakthanmanon T, Campana B, Mueen A, et al. Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping[J]. Acm Transactions on Knowledge Discovery from Data, 2013, 7(3):965-991.

[7] Cleuziou G, Moreno J G. Kernel methods for point symmetry-based clustering[J]. Pattern Recognition, 2015, 48:2812–2830.

[8] Keogh E J, Pazzani M J. Derivative Dynamic Time Warping[C]//SDM. 2001, 1: 5-7.

[9] Tan L N, Alwan A, Kossan G, et al. Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training dataa)[J]. The Journal of the Acoustical Society of America, 2015, 137(3): 1069-1080.

[10] Ratanamahatana C A, Keogh E. Making time-series classification more accurate using learned constraints[C]. SDM, 2004.

[11] Li C, Ma Z, Yao L, et al. Improvements on EMG-based handwriting recognition with DTW algorithm.[C]// Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference. 2013:2144-2147.

[12] Adwan S, Arof H. On improving Dynamic Time Warping for pattern matching[J]. Measurement, 2012, 45(6):1609–1620.

An optimization framework based on Kriging method with additive bridge function for variable-fidelity problem*

Peng Wang[1], Yang Li[1] and Chengshan Li
School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, P.R. China
wangpeng305@nwpu.edu.cn, 15202927067@139.com, 642961227@qq.com

***Abstract*—Variable-fidelity optimization (VFO), which utilizes the precise value of high-fidelity (HF) model and underlying trend of low-fidelity (LF) model, has solved many computationally expensive problems by simulation-based design. Though it has been developed rapidly in recent years, the simpler and cheaper ones are still needed. In this paper, a new optimization framework based on Kriging method with additive bridge function for variable-fidelity problem is proposed. The simple additive bridge function is taken to construct the primal HF model with Kriging method. With the local and global search strategies, the sample sets can be updated and the HF model be refreshed. It is worth mentioning that the fusion of them not only makes the method easy to implement, but also helps to find the optimal result much faster. In order to illustrate the ideas and features of the proposed optimization framework clearly, a mathematic example is presented in detail. Furthermore, another two problems are analyzed, including an engineering problem. The results show that the proposed optimization framework is feasible and effective, indicating it is suitable to solve complicated variable-fidelity problems.**

***Keywords*- VFO; additive bridge function; Kriging method; search strategies**

I. INTRODUCTION

Because of the complexity of modern problems, computationally expensive simulation models have been developed to acquire the complex mathematical models increasingly. Surrogate model, or Metamodel, is referred to as a technique that makes use of the sampled data to build an approximate model, which can predict the output at untried points in the design space and find the optimal result [1]. Due to this advantage, it has received increasing attention in different areas, including aerodynamic analysis [2], structural design [3], optimization [4], multiple criteria decision making (MCDM) [5] and Variable-Fidelity modeling (VFM) [6]. Even so, how to choose sample points, how to build surrogate models, and how to get the accurate solution are still key issues for surrogate modeling [7].

In traditional surrogate-based optimization (SBO), a surrogate model is built relied entirely on the expensive high-fidelity simulation. But for most of the problems, building an adequately accurate surrogate model is really a hard work, so variable-fidelity methods have been developed to address this kind of issue by incorporating both low-fidelity (LF) and high-fidelity (HF) models into one optimization framework [8]. Although the absolute values of cheaper LF data may be unbelievable, the underlying trend of it is available. Once a small number of HF samples are taken to correct the absolute values, the variable-fidelity method is considered feasible. Taking aircraft aerodynamic analysis as an example, the classical aerodynamics or linearized supersonic panel code is used to construct the LF model, while the HF simulation is realized by precise calculation such as the finite element based Euler/Navier–Stokes solver [9-12]. Though there have been many methods to construct the surrogate models, such as Kriging [13], polynomial response surfaces[14], support vector machines[15] and space mapping[16], how to manage different models of varying fidelity is still a key point need to be studied. In recent years, there have been some successfully methods implemented by researchers in different fields. For instance, [17] introduces a gradient-based approach to multi-fidelity optimization. [18] combine the direct Gradient-Enhanced Kriging (GEK) with a newly developed Generalized Hybrid Bridge Function (GHBF) to improve the efficiency and accuracy of the existing VFM approach. An investigation of aerodynamic models with varying fidelity is performed by [19].

In this paper, we present a new optimization framework based on Kriging method with additive bridge function for variable-fidelity problem. The rest of this paper is organized as follows. In Section II, the proposed optimization framework will be first presented. Then, a one-dimensional example is solved in considerable detail in Section III. In Section IV, another two practical problems are analyzed, including a two-dimensional problem (the Modified Branin Function) and an engineering problem (the design of AUV's shell). Conclusion is shown in Section V finally.

II. PROPOSED OPTIMIZATION FRAMEWORK WITH VARIABLE-FIDELITY METHOD

As shown in Fig.1, there are four basic stages in the proposed optimization framework with variable-fidelity method.

*This research was supported by the National Natural Science Foundation of China under Grant No. 51375389 and Fundamental Research Funds for the Central Universities under Grant No. 3102014JCQ01007.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.104

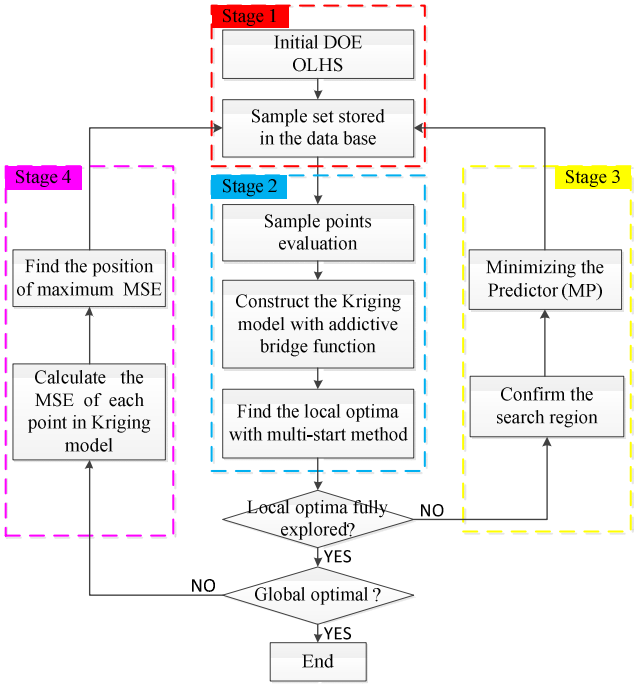


Fig.1. The proposed optimization framework with variable-fidelity method.

The first stage is sampling, applying the optimal Latin hypercube sampling (OLHS) [20] to get the sample sets in the experiment design [21]. Then, the surrogate models of both LF and HF will be constructed at the second stage, using simple "additive bridge function" [18] with Kriging method. The multi-start method [22], which is always combined with other local optimization methods, is used to explore the parameter space at the same stage. In order to make the models more accuracy, we take refreshing the sample sets into consideration. Both the local search strategy: Minimizing the Predictor (MP) [23] and the global search strategy: Maximizing the MSE are taken to help optimize the model. With the new points updated by two search strategies, the surrogate model can be much closer to the real model, helping us to find the global optimal result faster.

Though the methods we used are all existing ones, the combination adopts their good customs and gets rid of the bad ones. This not only makes the proposed method easy to implement, but also helps us to find the optimal result much faster, reducing both time and calculation a lot.

III. A NUMERICAL PROBLEM SOLVED TO EXPLAIN THE PROPOSED OPTIMIZATION FRAMEWORK

In this section, an example is used to illustrate the ideas and features of the proposed optimization framework. In this case, y_h represents the HF model and y_l is the LF one. The relationship between y_h and y_l is treated as unknown, which will be obtained by additive bridge function with Kriging method. The HF and LF models are shown as (1). The global solution of HF model is point x_{global} = 0.7572, and result y_{global} = -6.0207. The following process describes the basic steps of our optimization framework.

$$\begin{aligned} y_h &= (6x-2)^2 \sin(12x-4) \\ y_l &= 0.5y_h + 10(x-0.5) - 5 \\ x &\in [0,1] \end{aligned} \tag{1}$$

Step1 Sample and evaluate the points: One set with a small number of samples (0, 0.3, 0.5, 0.7, 1) for HF model and another significantly lager set (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) for LF model are calculated with HF method y_h and LF method y_l, respectively.

Step2 Construct the HF Kriging model with addictive bridge function: Based on the sample data, the response of the HF Kriging model $\hat{y}_h(x)$ is obtained by adding the addictive bridge function to the LF Kriging model $\hat{y}_l(x)$ approximately. $y_h(x)$, $y_l(x)$, $\hat{y}_h(x)$ and $\hat{y}_l(x)$ are plotted in Fig.2.

Step3 Search local optimal positions with multi-start method: The multi-start method is utilized to find the positions of all local minima. If all of them have been fully explored, the optimization will continue. Otherwise, the MP method is applied to find the local optimum further and add the minimum position to the sample sets.

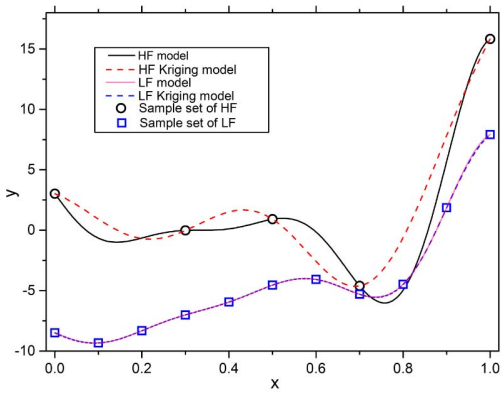


Fig. 2 The plots of HF, LF, HF Kriging model and LF Kriging model.

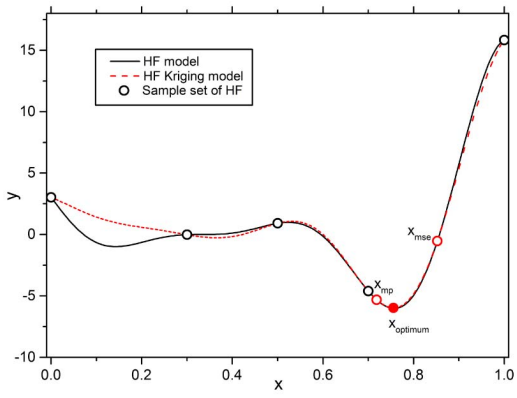


Fig. 3 The plots of HF Kriging model and the global optimal position

Step4 Find the position with maximum of the MSE for HF Kriging model: Because the HF Kriging model is constructed based on the combination of LF model and addictive bridge function, there must be some difference between it and the real high model. The larger MSE of the point is, the less accuracy it has. If we find the position of maximum MSE for HF Kriging model and add it to the point sets, the model will be refreshed and the average MSE will be decreased sharply.

Step5 Update the sample points and renew the HF Kriging model: The points getting in Step3 and Step4 are added to the sample sets in this step. This work ensures the model be closer to the real model, which will help us to acquire the global optimal position much faster. As shown in Fig. 3, the HF Kriging model is drastically improved by bringing in two special points.

Step6 Refinement: Iterative refinement is performed until a criterion for termination is fulfilled.

IV. EXAMPLE DEMONSTRATIONS

A. A Two-Dimensional Problem (the Modified Branin Function).

In order to test and verify the effectiveness of the proposed optimization framework, a modified Branin function is demonstrated in this section. By adding an additional small term, the original Branin function is modified that it has only one global minimum. The expressions of HF and LF models are given as follow (2). The ranges of x_1 and x_2 are normalized to 0 ~ 1 and the global theoretical solution of HF model is x_{global} = (0.0809, 1), y_{global}= -333.916.

$$y_h = y_{branin}(x_1, x_2) - 22.5x_2 \quad ,$$
$$y_l = y_{branin}(0.7x_1, 0.7x_2) - 15.75x_2 + 20(0.8 + x_1) - 80, \quad (2)$$
$$x_1 \in [-5,10], x_2 \in [0,15]$$

With the help of OLHS, the LF Kriging model is constructed with twenty points while the HF constructed with ten points. Four points are added by two strategies with the iterations continuing. The sample sets are updated and the HF Kriging model is refreshed. we calculated only fourteen points with HF model in this example. The optimal process is shown in Fig. 4 and Fig.5. The final optimal result is x= (0.0809, 1), y= -333.9160, which exactly matches with the global theoretical solution, confirming the effectiveness of the proposed optimization framework.

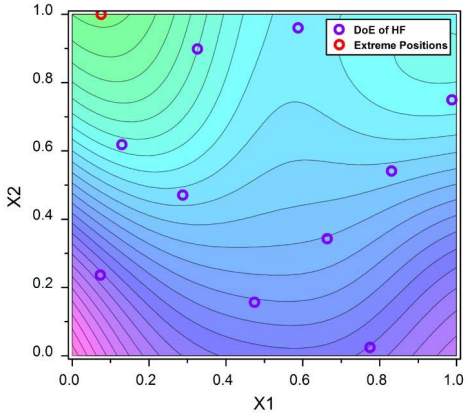


Fig. 4 Contours of the HF Kriging model and the position of first optimal solution.

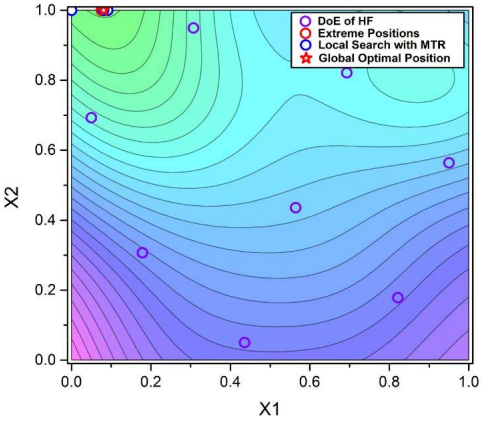


Fig. 5 Contours of the HF Kriging model and the position of final global optimal solution.

B. An Engineering Problem (the design of AUV's shell)

Autonomous Underwater Vehicles (AUVs) are unmanned, self-propelled robotic devices. The strength and stability of the shell structure must meet the requirement of large operational depth. Considering the middle section of torpedo-shaped AUV is more liable to failure, we only try to design this part. The ring-stiffened cylindrical shell structure with rectangle ribs is chosen and the structure style and design parameters of it are shown in Fig.6. Where t is the thickness of the shell, l donates the distance between the ribs, and t_2, l_2 represent the thickness and height of the ribs.

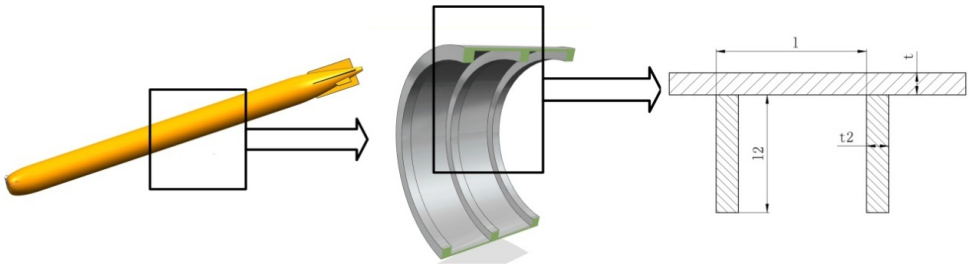


Fig. 6. The structure style and design parameters of shell.

In this problem, the diameter of the shell is 324mm and the depth of the water is 700m. The maximum von Mises stress needs to be less than 85% of the ultimate strength and the maximum critical load should be more than 120% of the calculating pressure. The goal of the optimization is to minimize the mass of the structure while satisfy the restrain condition, which is defined as follows:

$$\begin{aligned} &Max \quad M(t,n,t_2,l_2) \\ &st. \quad 4mm \le t \le 6mm \qquad 13 \le n \le 25 \\ &\quad 15mm \le t_2 \le 23mm \quad 12mm \le l_2 \le 16mm \\ &\quad 0 \le \sigma_{\max}(t,n,t_2,l_2) \le 0.85\sigma_s \\ &\quad 1.2P_j \le P_{cr}(t,n,t_2,l_2) \end{aligned}$$

Where σ_s refers to the ultimate strength of the Aluminium alloy and the and the water pressure P_j is calculated as $7.7MPa$.

Considering the problem has four design variables, we use OLHS to sample 40 points and we calculate all sample points with LF method and only 20 of them with HF method. As a result, the seventh optimization has gained a satisfactory result for the mass change:

$$\omega = \frac{|24.610 - 24.529|}{24.529} = 0.33\% < 1\%.$$

So $M(4.218 \quad 13 \quad 15.00 \quad 14.98)$ is treated as the ultimate result. The detailed optimize processes are shown in TABLE I. Where "MP" represents the minimum object of the predictor, "MMSE of $\sigma_{\max}$" and "MMSE of P_{cr}" mean the point with maximum MSE of $\sigma_{\max}$ model and P_{cr} model near to minimum object, respectively.

With the proposed optimization framework, we used 39 HF points and 59 LF points in all. The total calculation time is about 100 minutes, which is far less than the cost of just using a lot of HF points. According to the comparison of optimization processes, we can find that the calculation accuracy is greatly improved, which is usually reached by a lot of sample points. In conclusion, our proposed optimization framework has better performance in finding the optimum solution, saving both time and effort a lot.

TABLE I. THE DETAILED OPTIMIZE PROCESSES

	design variables				***objective***	***constraints***			
	t/mm	n	t_2/mm	l_2/mm	M/kg	$\sigma_{\max-H}$	P_{cr-H}	$\sigma_{\max-L}$	P_{cr-L}
The first Multi-start optimization									
MP	4.000	13	15.00	12.98	22.792	303.42	6.6955	266.15	7.6486
MMSE of $\sigma_{\max}$	4.437	14	15.88	12.25	25.037	267.54	7.3065	221.88	8.2151
MMSE of P_{cr}	4.000	13	15.00	12.00	22.320	300.79	5.8038	264.96	6.6851
The second Multi-start optimization									
MP	4.000	14	15.00	14.53	24.089	303.51	8.9020	238.83	10.129
MMSE of $\sigma_{\max}$	4.221	15	15.78	13.99	25.645	280.05	9.4695	224.48	10.654
MMSE of P_{cr}	4.000	13	15.00	16.00	24.225	312.19	10.143	270.76	11.378
The third Multi-start optimization									
MP	4.000	15	15.00	14.39	24.569	297.79	9.2492	233.78	10.514
MMSE of $\sigma_{\max}$	6.000	13	15.00	16.00	32.110	204.52	14.398	178.31	15.328
MMSE of P_{cr}	6.000	13	15.00	16.00	32.110	204.52	14.398	178.31	15.328
The fourth Multi-start optimization									
MP	4.000	14	16.88	14.36	24.969	302.13	9.421	235.63	10.685
MMSE of $\sigma_{\max}$	4.000	13	19.28	12.00	24.042	306.47	6.8177	258.63	7.8063
MMSE of P_{cr}	5.190	13	15.00	16.00	28.929	238.67	12.574	202.64	13.586
The fifth Multi-start optimization									
MP	4.282	13	15.00	14.80	24.778	285.61	9.1368	249.14	10.216
MMSE of $\sigma_{\max}$	6.000	13	15.00	12.00	30.231	194.68	8.7834	178.18	9.4133
MMSE of P_{cr}	6.000	13	15.00	12.00	30.231	194.68	8.7834	178.18	9.4133
The sixth Multi-start optimization									
MP	4.194	13	15.00	15.01	24.529	293.56	9.2346	250.85	10.35

MMSE of $\sigma_{\max}$	5.171	13	15.00	12.81	27.351	230.47	8.3091	201.09	9.0929
MMSE of P_{cr}	5.254	13	15.00	12.00	27.292	225.17	7.5774	202.51	8.2844
The seventh Multi-start optimization									
MP	4.218	13	15.00	14.98	24.610	291.13	9.2415	246.81	10.351
MMSE of $\sigma_{\max}$	6.000	13	15.00	13.54	30.961	197.23	10.674	177.39	11.398
MMSE of P_{cr}	4.000	13	23.00	16.00	28.460	310.27	13.646	250.44	15.357

V. CONCLUSIONS

As a recently developed method, VFO has solved many computationally expensive problems by taking full advantage of HF model and LF model in the optimization framework. Though it has a wide application now, the simpler and cheaper methods are still needed. In this paper, we proposed a new optimization framework based on Kriging method with additive bridge function for variable-fidelity problem. In order to improve the efficiency of optimization, we make fusion to some methods of mature, which adopts their good customs and gets rid of the bad ones. Three illustrative problems are successfully solved by the proposed optimization framework later, indicating it is suitable to solve the variable-fidelity problem. The main conclusions can be addressed as follows:

1) Expensive but less high-fidelity points, more but cheaper low-fidelity points and the simple idea of additive bridge function all make it easy for us to construct the approximate model. The combination of them reduces both the amount of calculation and the cost of time.
2) Considering the HF model got by simple additive bridge function may not so exact to the real model. For the local **s**earch strategy MP, by adding extreme positions of the HF model, we can find a right direction for the next search. Meanwhile, with the points of larger MSE infilling to the sample sets, which is the global strategy, the accuracy of the HF model can be greatly improved. It is obvious that two search strategies can make up the drawback caused by the primal construction of model, helping us to find the right direction of optimal result faster.

REFERENCES

[1] Nicolosi F, Della Vecchia P, Corcione S. Aerodynamic analysis and design of a twin engine commuter aircraft[C]//28th ICAS (International Council of Aeronautics and Astronautics) Conference. 2012: 23-28.

[2] Queipo N V, Haftka R T, Shyy W, et al. Surrogate-based analysis and optimization[J]. Progress in aerospace sciences, 2005, 41(1): 1-28.

[3] Lee K H. A robust structural design method using the Kriging model to define the probability of design success[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2010, 224(2): 379-388.

[4] Kleijnen J P C, van Beers W, Van Nieuwenhuyse I. Expected improvement in efficient global optimization through bootstrapped Kriging[J]. Journal of global optimization, 2012, 54(1): 59-73.

[5] Wang P, Meng P, Zhai J Y, et al. A hybrid method using experiment design and grey relational analysis for multiple criteria decision making problems[J]. Knowledge-Based Systems, 2013, 53: 100-107.

[6] Han Z H, Zimmermann R, Görtz S. A new coKriging method for variable-fidelity surrogate modeling of aerodynamic data[C]//48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, Orlando, Florida. 2010: 4-7.

[7] Koziel S, Ogurtsov S. Surrogate-based optimization[M]//Antenna Design by Simulation-Driven Optimization. Springer International Publishing, 2014: 13-24.

[8] Y. Kuya, K. Takeda, X. Zhang, A.I. J. Forrester, Multifidelity surrogate modeling of experimental and computational aerodynamic data sets, AIAA journal, 49 (2011) 289-298

[9] Xiong Y, Chen W, Tsui K L. A new variable-fidelity optimization framework based on model fusion and objective-oriented sequential sampling[J]. Journal of Mechanical Design, 2008, 130(11): 111401.

[10] Slotnick J, Clark R W, Friedman D M, et al. Computational Aerodynamic Analysis for the Formation Flight for Aerodynamic Benefit Program[C]//AIAA Science and Technology Forum and Exposition. 2014.

[11] Franco A, Valera D L, Pena A, et al. Aerodynamic analysis and CFD simulation of several cellulose evaporative cooling pads used in Mediterranean greenhouses[J]. Computers and Electronics in Agriculture, 2011, 76(2): 218-230.

[12] Beechook A, Wang J. Aerodynamic analysis of variable cant angle winglets for improved aircraft performance[C]//Automation and Computing (ICAC), 2013 19th International Conference on. IEEE, 2013: 1-6.

[13] Ankenman B, Nelson B L, Staum J. Stochastic Kriging for simulation metamodeling[J]. Operations research, 2010, 58(2): 371-382.

[14] Venkatesh V, Goyal S. Expectation disconfirmation and technology adoption: polynomial modeling and response surface analysis[J]. MIS quarterly, 2010, 34(2): 281-303.

[15] Murty M N, Devi V S. Support Vector Machines[M]//Pattern Recognition. Springer London, 2011: 147-187.

[16] Koziel S, Cheng Q S, Bandler J W. Space mapping[J]. Microwave Magazine, IEEE, 2008, 9(6): 105-122

[17] Balabanov V, Venter G. Multi-fidelity optimization with high-fidelity analysis and low-fidelity gradients[C]//10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. 2004, 4459.

[18] Han Z H, Görtz S, Zimmermann R. Improving variable-fidelity surrogate modeling via gradient-enhanced Kriging and a generalized hybrid bridge function[J]. Aerospace Science and Technology, 2013, 25(1): 177-189.

[19] Wilke G. Variable fidelity optimization of required power of rotor blades: Investigation of aerodynamic models and their application[C]//38th European Rotorcraft Forum. 2012.

[20] Kucherenko S, Albrecht D, Saltelli A. Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques[J]. 2011.

[21] Kirk R E. Experimental design[M]. John Wiley & Sons, Inc., 1982.

[22] Martí R. Multi-start methods[M]//Handbook of metaheuristics. Springer US, 2003: 355-368.

[23] Forrester A I J, Keane A J. Recent advances in surrogate-based optimization[J]. Progress in Aerospace Sciences, 2009, 45(1): 50-79.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Surrogate-Based Optimization for Autonomous Underwater Vehicle's Shell Design

Huachao Dong, Baowei Song, Peng Wang
School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an 710072, P.R.China
E-mail: wangpeng305@nwpu.edu.cn

***Abstract*—According to the AUVs' mission requirement of large operational depth, this paper provides a kind of ring-stiffened multiple intersecting spherical shell (RSMISS) for AUVs. In order to get the performance of the new shell, a surrogate-based optimization (SBO) method is proposed. The SBO method is employed to explore the kriging-based model. Simultaneously, a multi-start optimization approach is repeated to do a global search. Furthermore, an enhanced "Minimizing the Predictor" (EMP) method is presented as the local infill criteria which can dynamically adjust the search direction and the length of sampling radius. Ultimately, the proposed SBO method is applied to look for the agreeable solution of the RSMISS. The results show that EMP can find the satisfactory solution more quickly with smaller computational cost.**

Keywords- Autonomous underwater vehicle; Multiple intersecting spherical shell; Surrogate-based optimization; Minimizing the Predictor; Kriging

I. INTRODUCTION

Autonomous Underwater Vehicles (AUVs) are unmanned, self-propelled robotic devices, which have an increasingly ubiquitous role in the scientific, commercial and military field in recent years [1]. The traditional design of AUVs' shell structure is based on the empirical formulas of the torpedo shell [2]. The main type of the torpedo shell is ring-stiffened cylindrical shells (RSCSs) However, with the depth of the marine increasing, RSCSs can't provide the permissible stress and better stability [3]. Although, advanced composite material can enhance the performance of AUVs' shell, the cost of production will be added enormously. Furthermore, the expenses of the compressive strength testing are also quite expensive. Hence, in the conceptual design stage, changing AUVs' shell structure by Computer Aided Engineering (CAE) to improve its strength and stability becomes a significant means [4].

In this paper, according to the actual mission requirement of large operational depth, one kind of ring-stiffened multiple intersecting spherical shell (RSMISS) appropriate for AUVs was provided. In the process of conceptual design, in order to offer the optimal consequence for the subsequent detailed design, surrogate-based optimization (SBO) is employed.

SBO proves to be an efficient and feasible approach for simulation-based conceptual design, which can greatly reduce computational cost [5]. Generally, SBO includes three phases: infill criterion, constructing surrogate models (SM) and optimizing surrogate models. Infill criterion is a plan of sampling, which provides a set of suitable sample points to construct an accurate surrogate model [6]. There exist several frequently-used construction methods that are the response surface method (RS) [7], the kriging interpolation method [8], and the radial basis function (RBF) method [9], respectively. Kriging treats the system response as a realization of a stochastic process to predict the nonlinear problem better. "Minimizing the Predictor" (MP) is a kind of infill criterion which minimizes the predictor and runs several true function evaluations within the neighboring region around the optimal value [7]. However, this method depends on the accuracy of the initial surrogate model and can't solve multimodal engineering problems. What's more, if the optimal solution of the predictor is far away from the true optimal point, it will cost lots of expensive evaluations.

In this paper, according to the characteristics of RSMISSs, the kriging-based model is used and a SBO process is proposed. The process involves a multi-start optimization method and a local infill criterion called "Enhanced MP" (EMP). The presented EMP method explores the unknown region by a management framework of a SM, which can dynamically adjust the search direction and the length of sampling radius to decrease the computational cost. Ultimately, the proposed SBO method is implemented to obtain the optimal design of the presented AUVs' RSMISSs. Through comparative analysis, RSMISSs performs better than the traditional RSCSs on AUVs.

II. PROPOSED SURROGATE-BASED OPTIMIZATION TECHNOLOGY

First of all, several sample points will be obtained by design of experiment (DOE) to construct the kriging-based SM [10]. After the SM is initialized, a multi-start optimization method is used, which utilized a local SQP optimizer from different starting points to look for local optimal solutions of the SM. Simultaneously, these local optimal solution will be evaluated and added to the previous sample set. This process is repeated until a stop condition is satisfied. With evaluations increasing, the previous local optimal positions may move to the true local optimal location or just disappear [11]. This stop condition confirms the current local optimal solution of the SM is near the true one. After the local optimal solutions of the SM obtained, the local search begins to run. In order to obtain a better design, an Enhanced "Minimizing the Predictor" method (EMP) is proposed to fully explore the local region with fewer simulating calculations. Fig.1 shows the optimization process.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.105

In the EMP method, the best local optimal solution of the SM is regarded as the initial point $\boldsymbol{x}_0$ and meanwhile $f(\boldsymbol{x}_0)$ is evaluated to refit the SM. Since kriging-based SMs have a better approximation performance in the neighborhood of sample points, we set a search radius to search the next point $\boldsymbol{x}_1$. Users define the initial search radius $\boldsymbol{\delta}_0$. The next points are obtained by optimizing the kriging model in the region $\boldsymbol{x}_0 \pm \boldsymbol{\delta}_0$. The new SM will be reconstructed by new points. $\boldsymbol{\delta}_1$ will also be updated as follows. Equation (1) is used to evaluate the performance of the prediction.

$$r = \frac{f(\boldsymbol{x}_{m-1}) - f(\boldsymbol{x}_m)}{f(\boldsymbol{x}_{m-1}) - \hat{y}(\boldsymbol{x}_m)} \tag{1}$$

Where, $f(\boldsymbol{x}_{m-1})$ and $f(\boldsymbol{x}_m)$ are the evaluations at the prior and current points. $\hat{y}(\boldsymbol{x}_m)$ is the response value of SM at the current point $\boldsymbol{x}_m$. The new search radius can be calculated through the formula (2).

$$\boldsymbol{\delta}_m = \begin{cases} \{c_1 \,|\, x_m^1 - x_{m-1}^1 \,|, c_1 \,|\, x_m^2 - x_{m-1}^2 \,|, \cdots c_1 \,|\, x_m^n - x_{m-1}^n \,|\} \; if \; r < r_1 \\ \{c_2 \,|\, x_m^1 - x_{m-1}^1 \,|, c_2 \,|\, x_m^2 - x_{m-1}^2 \,|, \cdots c_2 \,|\, x_m^n - x_{m-1}^n \,|\} \; if \; r > r_2 \\ \{|\, x_m^1 - x_{m-1}^1 \,|, |\, x_m^2 - x_{m-1}^2 \,|, \cdots |\, x_m^n - x_{m-1}^n \,|\} \quad otherwise. \end{cases} \tag{2}$$

The parameter n denotes the dimension of the design vector. The coefficients c_1 and c_2 are known as the degree of expansion and shrinkage for the new search region. The parameters r_1 and r_2 determine the search region when to expand or shrink [7]. Although local search is implemented in a trust region, the algorithm may still lose the right direction. This is because with the search region expanding, the approximation performance of the kriging-based SM will deteriorate. Here a parameter ε is set as follows.

$$\varepsilon = \begin{cases} 0 \;\; if \;\; f(\boldsymbol{x}_m) > f(\boldsymbol{x}_{m-1}) \bigcap f(\boldsymbol{x}_{m-1}) > f(\boldsymbol{x}_{m-2}) \\ 1 \quad otherwise \end{cases} \tag{3}$$

ε is defined as 0 or 1. (For minimization problems) When ε is 1, it means that the next point performs better than the prior one or the next point may have chance to change the direction into the right way and iterations go on along the direction based on the trust-region framework. On the contrary, when ε is 0, it means the next point loses the right direction for successive two times and d+1 points will be obtained by Optimal Latin Hypercube Sampling (d is the dimension of $\boldsymbol{x}$) in the vicinity of $\boldsymbol{x}_{m-1}$ to improve the approximation effect of the SM there [7]. The radius of re-sampled area around $\boldsymbol{x}_{m-2}$ is $(|\, x_m^1 - x_{m-1}^1 \,|, |\, x_m^2 - x_{m-1}^2 \,|, \cdots |\, x_m^n - x_{m-1}^n \,|)$. Here, n denotes the dimension. After the right direction is achieved, the algorithm continues until the convergence. In this paper, the condition of convergence is as follows

$$\begin{cases} g(\boldsymbol{x}_m) \geq 0 \\ |\, f(\boldsymbol{x}_m) - y_{goal} \,| \leq 0.5\% \cdot |\, y_{goal} \,| \end{cases} \tag{4}$$

Here, y_{goal} is the expected value of the true black-box problem. The formula $g(\boldsymbol{x}_m) \geq 0$ denotes that the constraint need be satisfied.

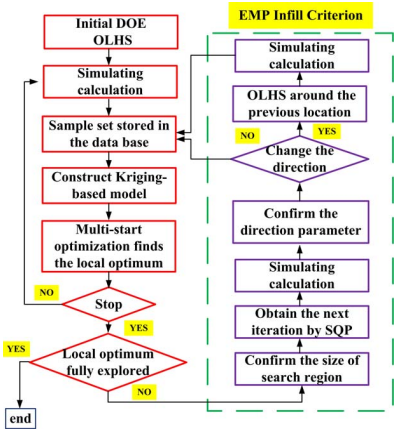


Figure 1. The proposed surrogate-based optimization process

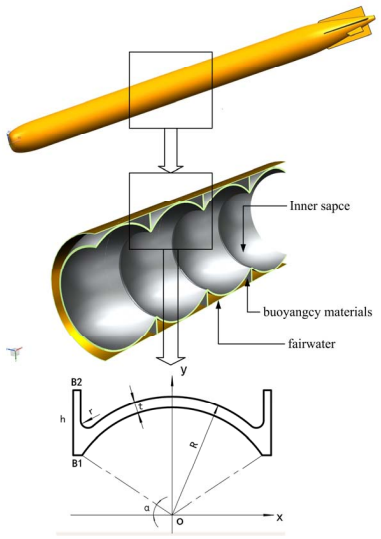


Figure 2. The ring-stiffened multiple intersecting spherical shell of AUVs

III. RING-STIFFENED MIS SHELL FOR AUVS

The traditional shell of the torpedo-shaped AUV is the RSCS which is commonly designed under the condition of 500m's depth. As the depth of water increases, the mass will get larger and the shell will be more unstable [2]. Therefore, a kind of new shell structure (RSMISS) is proposed. The multiple intersecting spherical shell has proven to have high strength and high stability [12]. However, no one tries to utilize it on AUVs. In order to obtain more spacious internal space of AUVs, ring-shaped ribs are set up around the outer hull as Fig. 2 shows. According to the cylindrical characteristic of torpedo-shaped AUVs, the RSMISS is provided and analyzed on AUVs.

The structural style and design parameters of RSMISSs are shown in Fig. 2. Among these parameters, R defined as a constant denotes the radius of shell. B_1, B_2, h, are width and height of main and auxiliary ribs, respectively. Since the auxiliary rib is responsible for the safety of bolted connection between two spherical shells and auxiliary rib can be regarded as the extension of the main rib, the parameters B_2 will be defined equal to B_1 in the subsequent optimization process. The distance between the top of the rib and the horizontal axis is set equal to R. At last four design variables

are confirmed as follows: the thickness of shell t, the radius of face blend r, the radian of shell α and the width of main rib B_1.

In this paper, R is 162 mm, the depth of water is 1000 m and four multiple intersecting spheres are analyzed. The mechanical properties of Aluminum alloy are shown in TABLE I, which are used in the finite element analysis (FEA). Buoyancy-weight ratio (BG), maximum von Mises stress ($\sigma_{\max}$) and maximum critical load of the RSMISS (P_{cr}) are calculated as response values, respectively.

IV. SURROGATE-BASED OPTIMIZATION ON AUVS' RSMISS

As our previous discussion, there are four design variables and three response values in this optimization problem. The objective is maximizing the buoyancy-weight ratio. Two security coefficients are proposed in constraints. The maximum von Mises stress needs to be less than 85% of the ultimate strength. Meanwhile, the maximum critical load should be more than 120% of the calculating pressure. In order to meet the requirement of the inner space and processing feasibility, the specific optimization expression is defined as follows:

TABLE I. MECHANICAL PROPERTIES OF ALUMINUM ALLOY

Properties	Aluminum Alloy
Young's Modulus	70Gpa
Desnsity	2800 kg/m^3
Ultimate Strength	4.5Mpa
Poisson's Ratio	0.3

$$\begin{aligned} &Max\ BG(t,\ B_1, r,\ \alpha) \\ &st.\quad 2.5mm \le t \le 6mm \quad 3mm \le B_1 \le 7mm \\ &\qquad 3mm \le r \le 7mm \quad 0.6981rad \le \alpha \le 1.3089rad \\ &\qquad 0 \le \sigma_{\max}(t,\ B_1, r,\ \alpha) \le 0.85\sigma_s \\ &\qquad 1.2P_j \le Pcr(t,\ B_1, r,\ \alpha) \end{aligned} \quad (5)$$

Firstly, $3n$+2 (It means 14 points) sample points are obtained by DOE and evaluated by FEA. Three kriging-based models are constructed through these sample points, which can predict buoyancy-weight ratio, maximum von Mises stress and maximum critical load of the RSMISS. Since the initial DOE just provides 14 sample points, the accuracy of the predictor is low and some unreal local optimal locations appear. In this paper, the multi-start optimization approach is employed to acquire these local optimal locations of kriging-based models and the predictor gets updated. The stop condition mentioned in Section II is that when two adjacent solutions are closer and the feasible solution appears, the multi-start optimization stops. In this example, five iterations are carried out and a local area is confirmed, which is a neighboring region around $(3.62, 4.84, 6.86, 0.7887)$.

TABLE II shows intermediate results about the multi-start optimization process. From TABLE II, a discovery can be obtained that the number and location of local optimal solutions change as the iteration runs. Meanwhile, the overall accuracy of the SMs improves. After the fifth iteration, the feasible solution (3.62, 4.84, 6.86, 0.7887) is found that is closer to the solution (3.76, 4.40, 7, 0.8308) from the forth iteration.

TABLE II. THE LOCAL OPTIMAL SOLUTIONS OF THE KRIGING-BASED PREDICTOR AND RESPONSE VALUES SOLVED BY FINITE ELEMENT ANALYSIS

The local optimal solutions	**design variables**				**objective**	**constraints**		**feasiblility**
	t/mm	B_1/mm	r/mm	a/rad	BG	σ_{max}/MPa	Pcr/MPa	*Yes/No*
After the first multi-start optimization	3.67	5.76	7.00	0.698	4.248	340.40	56.429	*Yes*
	6.00	7.00	7.00	0.698	2.906	237.23	124.056	*Yes*
	6.00	4.16	3.00	0.698	3.242	364.19	89.311	*Yes*
	6.00	4.26	7.00	0.698	3.156	262.25	95.072	*Yes*
	6.00	7.00	7.00	1.309	3.745	318.50	13.476	*No*
After the second multi-start optimization	2.50	5.41	7.00	0.763	5.746	473.91	26.394	*No*
	2.50	6.12	7.00	0.790	5.639	465.29	26.452	*No*
	6.00	4.23	7.00	1.309	3.798	334.28	13.396	*No*
	2.50	6.37	6.61	0.757	5.457	433.88	32.575	*No*
	5.34	4.92	7.00	1.309	4.206	368.27	10.798	*No*
After the third multi-start optimization	2.86	5.59	5.52	0.728	5.212	420.38	39.646	*No*
	4.85	5.40	3.00	1.031	4.283	477.43	11.295	*No*
	5.34	4.11	3.00	1.178	4.227	487.96	9.008	*No*
Afterthe forth multi-start optimization	3.76	4.40	7.00	0.830	4.764	391.97	32.542	*No*
	2.50	6.71	4.95	0.836	5.804	439.43	35.437	*No*
	5.62	7.00	3.35	1.132	3.824	408.84	9.497	*No*
After the fifth multi-start optimization	3.62	4.84	6.86	0.788	4.708	374.85	44.608	*Yes*
	2.50	7.00	5.56	0.728	5.241	426.82	34.762	*No*
	6.00	5.03	7.00	1.257	3.745	343.92	12.225	*No*
	5.09	4.39	7.00	1.214	4.320	380.18	9.551	*No*
Best Solution	3.62	4.84	6.86	0.788	4.708	374.85	44.587	*Yes*

In order to further improve the performance of the shell structure, EMP is employed for local search. Through these samples from DOE and the multi-start process, designers need to provide an expected objective value y_{goal} as Equation (4) advises. In this paper, y_{goal} is set equal to 4.85. Simultaneously, the initial point is firstly defined as $\boldsymbol{x}_0 = (3.62, 4.84, 6.86, 0.7887)$ and the initial search region is set as $\delta_0 = (0.14, 0.45, 0.15, 0.05)$. In this example, the objective function $f(\boldsymbol{x})$ was modified by a penalty function method to implement Equations (1) and (3) [13]. On the basis of the optimization process proposed by Fig.1, the EMP method runs and the procedure's results are shown in TABLE III. In the process of EMP, after four iterations the right direction is lost and five sample points are added to update the Kriging-models. Eventually, the satisfactory solution is obtained with ten evaluations by FEA. On the other hand, as TABLE IV suggests, the MP method runs with eleven evaluations, while a better solution is not obtained.

The simulation results which involve the distribution of von-Mises stress and the total deformation by FEA are displayed in Figure 3. Figure 3 (a) (b) shows the best results obtained after the multi-start optimization process. Figure 3 (c) (d) shows the best results obtained after the local search by EMP. For the sake of a better visuality, the figures of total deformation are magnified by a factor of twenty eight. As TABLE II shows, the initial kriging model is optimized for the first time and the best solution $(3.67, 5.76, 7, 0.6981)$ is acquired. The obtained solution is feasible and its buoyancy-weight ratio is 4.25. After the multi-start optimization process is implemented, the best performance is improved from 4.25 to 4.71 and the rate of increase is 10.82%. In addition, TABLE III indicates that the local infill criterion (EMP) improves the performance from 4.71 to 4.848, the farther growth rate is 2.9%. Simultaneously, the comparison of TABLE III and TABLE IV suggests that the EMP method can find the agreeable solution more quickly than MP. In conclusion, the proposed SBO process effectively improves the performance of RSMISSs at the concept design stage.

V. CONCLUSION

According to the torpedo-shaped AUVs' characteristics, a kind of RSMISS is proposed in this paper. In order to explore the black-box model, a SBO process is provided to acquire the optimal design. Based on FEA, the SBO process finds the satisfactory solution with fewer simulating evaluations. In addition, the RSMISS performs better and may be more suitable for AUVs in the future. The main conclusion can be addressed as follows:

(1) The RSMISS can provide larger buoyancy-weight ra-

TABLE III. THE LOCAL SEARCH PROCESS WITH THE EMP INFILL CRITERIA

The EMP iteration procedures	**design variables**				**objective**	**constraints**		**feasiblility**
	t/mm	B_1/mm	*r*/mm	*a*/rad	*BG*	σ_{max}/MPa	*Pcr*/MPa	*Yes/No*
The first iteration	3.5643	5.0739	6.7113	0.8254	4.821	386.40	36.070	*No*
The second iteration	3.4821	4.7505	6.8366	0.7621	4.791	376.39	50.028	*Yes*
The third iteration	3.3588	4.3853	6.7970	0.7282	4.910	387.94	50.054	*No*
The forth iteration	3.3943	4.7290	6.7763	0.7623	4.886	388.59	48.869	*No*
Confirm the right search direction by OHLS	3.4548	4.7655	6.8585	0.7620	4.813	391.52	49.817	*No*
	3.5093	4.7571	6.8187	0.7622	4.765	378.79	50.266	*Yes*
	3.4821	4.7402	6.7988	0.7620	4.795	387.93	49.961	*No*
	3.5365	4.7486	6.8785	0.7621	4.737	387.35	50.611	*No*
	3.4310	4.7350	6.8410	0.7622	4.844	395.19	49.380	*No*
The fifth iteration	3.4276	4.7318	6.8386	0.7622	4.848	382.37	49.417	*Yes*
Optimal Solution	3.4276	4.7318	6.8386	0.7622	4.848	382.37	49.417	*Yes*

TABLE IV. THE LOCAL SEARCH PROCESS WITH THE MP INFILL CRITERIA

The MP iteration procedures	**design variables**				**objective**	**constraints**		**feasiblility**
	t/mm	B_1/mm	*r*/mm	*a*/rad	*BG*	σ_{max}/MPa	*Pcr*/MPa	*Yes/No*
First OHLS to update the Kriging-based model	3.4821	5.2923	6.8556	0.8137	4.816	374.72	39.546	*Yes*
	3.5521	4.3923	6.9278	0.7637	4.802	399.24	48.408	*No*
	3.6921	5.0673	6.7834	0.7387	4.479	352.37	56.726	*Yes*
	3.6221	4.6173	6.7113	0.8387	4.893	394.18	31.355	*No*
	3.7621	4.8423	7.0000	0.7887	4.576	379.32	45.446	*Yes*
Optimal solution	3.3548	5.2895	6.7000	0.8118	4.945	399.93	39.344	*No*
Second OHLS to update the Kriging-based model	3.4248	5.5145	6.8500	0.7868	4.750	367.03	47.239	*Yes*
	3.4948	5.2895	6.7000	0.8618	4.950	398.08	29.642	*No*
	3.2848	5.7395	6.6250	0.8468	5.032	429.70	33.171	*No*
	3.3548	5.0645	6.5500	0.7618	4.860	393.82	49.028	*No*
	3.2148	4.8395	6.7750	0.8118	5.193	408.69	36.491	*No*
Optimal solution	3.3001	5.1456	6.8471	0.763	4.884	384.25	48.1872	*No*

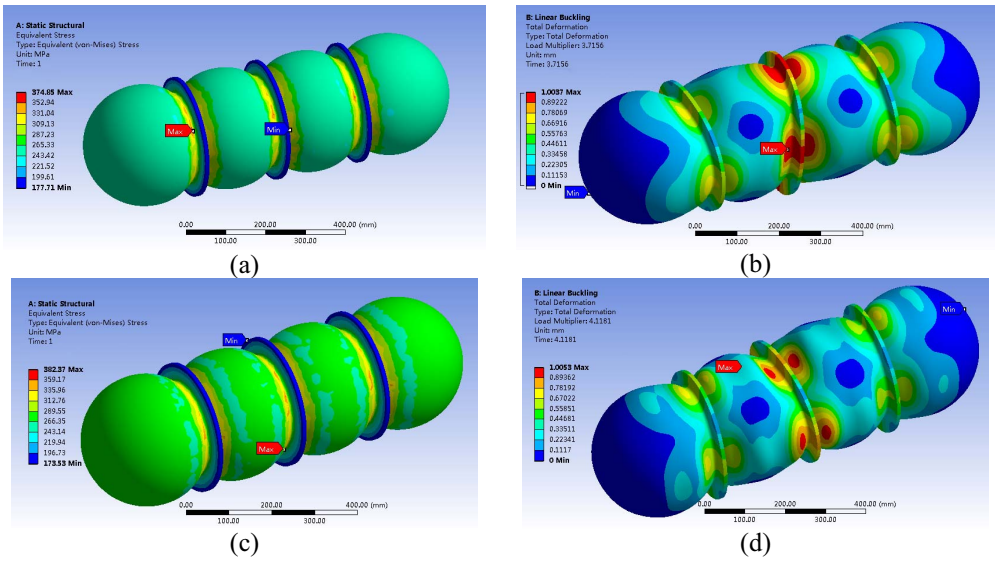


Figure 3. The distribution of von Mises stress and the total deformation at the RSMISS's different solutions .

tio and higher loading capacity (better stability).

(2) The proposed SBO process provides a novel idea on global and local exploration of kriging-based models. The multi-start optimization approach is utilized to do a global search. The presented EMP method is employed for local search. This is a novel approach for the black-box engineering problem.

(3) According to the defects of the MP method, an enhanced MP (EMP) method is presented. The EMP method can dynamically adjust its search direction and the length of sampling radius in the process of local search. The results show that EMP can find the satisfactory solution with fewer simulating evaluations than MP.

ACKNOWLEDGMENT

The author is grateful to the editor and the anonymous referees for their insightful comments. This research was supported by the National Natural Science Foundation of China (Grant No. 51375389).

REFERENCES

[1] Storkersen, N., Kristensen, J., Indreeide, A., Seim, J. and Glancy, T., "Hugin—UUV for seabed surveying." Sea Technology, vol. 39, no.2, pp.99-104,1998.

[2] Huachao Dong, Baowei Song and Peng Wang, "Multi-objective Optimal Design of Automatic Underwater Vehicle Shell Structure," Acta Armamentarii, vol.35, no.3, pp. 392-397,2014.

[3] Baowei Song, Pengfei Cheng and Yonghui Cao, "Strength and Stability of Multiple Intersecting Spheres for Pressure Hull," Computer Simulation, vol.30, no.2, pp.38-41, 2013.

[4] Huijiang He and Nan Li, "Parametric Modeling of Torpedo Shell Structure Based on APDL," Torpedo Technology, Vol.18, No.4,pp. 246-248,2010.

[5] Koziel S.and Ogurtsov S., "Surrogate-Based Optimization", Antenna Design by Simulation-Driven Optimization. Springer International Publishing, pp. 13-24,2014.

[6] Picheny V, Wagner T and Ginsbourger D. "A benchmark of kriging-based infill criteria for noisy optimization". Structural and Multidisciplinary Optimization, vol.48, no.3, pp.607-626, 2013.

[7] Forrester A. I. J.and Keane A. J., "Recent advances in surrogate-based optimization," Progress in Aerospace Sciences, vol.45, no.1, pp.50-79, 2009.

[8] Stein M. L., "Interpolation of spatial data: some theory for kriging," Springer, 1999.

[9] Park J. and Sandberg I. W., "Universal approximation using radial-basis-function networks," Neural computation, vol. 3, no.2, pp. 246-257, 1991.

[10] Chen V. C. P., Tsui K. L., Barton R. R., et al., "A review of design and modeling in computer experiments," Handbook of Statistics, vol.22, pp. 231-261, 2003.

[11] Regis R. G. and Shoemaker C. A. A., "quasi-multi start framework for global optimization of expensive functions using response surface models," Journal of Global Optimization, vol.56, no.4, pp.1719-1753, 2013.

[12] Cho-Chung Liang, Sheau-Wen Shiah, Chan-Yung Jen and Hung-Wen Chen, "Optimum design of multiple intersecting spheres deep-submerged pressure hull," Ocean Engineering, vol.31, pp. 177-199, 2004

[13] Di Pillo G. and Grippo L., "An exact penalty function method with global convergence properties for nonlinear programming problems," Mathematical Programming, vol.36, no.1, pp. 1-18, 1986.

The falling range prediction model of lost plane

Kaiyin Zhang Min Lan
Wuhan University of Technology, Wuhan, 430070
e-mail: iyguang@whut.edu.cn

Yan Huang Yuguang Li
Wuhan University of Technology, Wuhan, 430070
e-mail: Liyuguang@qq.com

Abstract: It is very important for searcher to predict the lost plane falling rang. The motion equation of lost plane is built with additional mass and random force, and solved by using difference method abd random number, the trajectory is calculated by using integral method, the lost plane falling rang is determined by using probability distribution contour map

Keywords-Aircraft Crash; Searchin;

Euler interpolation; Falling Probability Contour Plots.

I. INTRODUCTION

.MH370 still did not find. from crash until now[1]. Problem need to build a generic mathematical model that could assist "searchers" in planning a useful search for a lost plane feared to have lost in open water such as the Atlantic, Pacific, Indian, Southern, or Arctic Ocean while flying from Point A to Point B[2].

Aircraft after the crash, a complete search and rescue work is: first determined the site of aircraft crash, and then to find survivors and the implementation of the rescue process[3].When the aircraft crashed near the airport or on the runway at the airport, there is not search problems, but mainly to the rescue. But when the aircraft crashed in the mountains, desert, ocean, to determining the search area is particularly important[4]. If aircraft lost contact and communication in the absence of any of the sign, or too late to report their exact location has crashed, or complex terrain, bad weather, no witnesses report, make sure the search area and accurate work quite difficult[5]. At this time, it needs to identify several different areas to search. Lack in the number of personnel and aircraft, it is easy to miss the best time for rescue. So that, to determine search area has great significance to save lives and property in the limited time..

II. HULL DEFORMATION PREDICTION MODEL FOR THE SEARCH AREA

.After an aircraft accident, a complete search and rescue work is from determining aircraft crash site to rescue the survivors. The most common method of determining the search area is as follows: The search area is a circle. The center is the location of the aircraft reported finally, Search radius is the maximum glide distance when the fuel used up. However, this search area is too broad to miss the best time for rescue. So it’ s is very important to determine the search area according to the different search aircraft, the weather and the sea state when the aircraft lost contact.

Modern aircraft are equipped with ELBA （emergency location beacon-aircraft）, when the aircraft can’ t keep flying, it can transmit signal timely and indicate the location of the aircraft. So we think that the basic model has the clear location of the aircraft when it lost

Summary: we make the following assumptions about model.

Aircraft lost contact instantly lose power force;

Let sea level on the plane XOY , Z-axis is vertically upward, projection of original falling point at sea level is set as coordinate origin;

When the aircraft falls into water, part of the kinetic energy was lost, the amount of losses was determined by the state of aircraft . This article assumes that the aircraft the speed lost n% during the falling process, and n = 50.

When the plane lost contact at an altitude of H from sea level, and instantly lost power and started to fall. As Fig 1 shows that stalling spatial coordinates of the aircraft is (0,0, H).

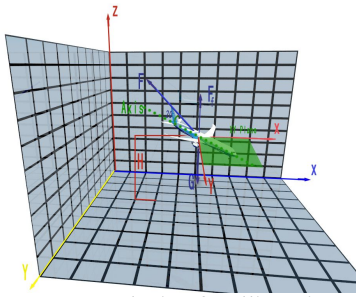


Fig.1 Force Analysis of stalling aircraft

Force Analysis of stalling aircraft is shown above. Let aircraft quality be m, then G= （0,0,mg） T, Fluid force $F=\lambda(F_L-F_r)-F_D$

Where λ denotes stalling state parameter; λ =1 predicts that aircraft keeps normal flight, λ =0 predicts that aircraft can't work without lift. Where FL is lift , Fr is induced drag, FD is the amount of shape resistance and friction..

During the falling process,aircraft is affected by,and.interfered by,motion trajectory of the aircraft is.space trajectory of the aircraft barycenter is given by

$$S+S_R=(x,y,z)^T+(x_R,y_R,z_R)^T \qquad (1.1)$$

V=(Vx,Vy,Vz) T=Fluid velocity;v=(vx,vy,vz) T = relative velocity between aircraft and fluid; According to Newton's second law,Random motion equation of the disappearing aircraft are detailed as follows[6]:

$$\begin{aligned} &m(\ddot{S}+\ddot{S}_R)+M(\ddot{S}+\ddot{S}_R)=F+G-F_f+F_R \\ &S(0)=S_0=(0,0,-H)^T \\ &\dot{S}(0)=\dot{S}_0=(v_x(0),v_y(0),v_z(0))^T \end{aligned} \qquad (1.2)$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.106

$$S_R(0)=0 \quad \dot{S}_R(0)=0$$

Where t is time,when t=0,aircraft is falling without power.S(0) denotes the initial position vector, represents the initial velocity, M represents the additional mass matrix.

F_L is lift, the calculating formula is：

$$|F_L|=\frac{1}{2}\rho V^2 C_L A,\ C_L=1.1488(\alpha+\theta)$$ In the expression, C_L is lift coefficient for the aircraft. $\frac{1}{2}\rho V^2$ is the plane flew operation pressure; A is the area of the wing. is the angle of attack, is the angle of the plane to flow. For Fr drag, the calculating formula is. is the sum of the shape resistance and the friction resistance. the calculating formula is：$|F_r|=F_L\cos(\alpha)$. In the expression,

$$|F_{Df}|=\frac{1}{2}\rho V^2 C_{Df} A,\ C_{Df}=\frac{0.074}{R_c^{0.2}},\ R_c=\frac{VC}{\nu}$$

C_{Df} is lift coefficient for the aircraft. $\frac{1}{2}\rho V^2$ is the plane flew operation pressure; A is the area of the wing. v is the fluid viscosity coefficient, air dynamic viscosity coefficient is 1.5*10E-5 m2/s, ,Dynamic viscosity coefficient of the water is 1.003*10E-6 m2/s.Above all, , is the fluid density; When the Angle of attack is zero, the projection area of the wings on xoy plane is A; Set up for angle of the wing and the plane. is the angle of the plane to flow. The sum of the both are $\alpha=\alpha_0+\theta$.

In the air, the air density ρ varies with height, its formula is:. In the expression, ρ 0 is the air mass density when the centigrade is zero at sea level, ρ 0=1.292kg/m^3。Tc=0.0065/273(1/m), In the sea, the density of the water is . ρ wdoes not change along with the water depth H.

The added mass matrix M general expressions for the:

$$M=\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix} \quad (1.3)$$

In the expression, , $m_{ij}=\rho\iint_A \phi_j n_i dA$ i,j=1,2,3. A is surface area of the aircraft, dA is the yuan area of the micro plane surface， ϕ_j is the velocity potential function of j direction unit of speed。for the plane surface vector in the direction of the i component.

In the air, the plane buoyancy approximate is neglected , namely.. $F_{fA}=0$. In the water, the buoyancy is equal to the weight of water replaced by the plane. then $F_{fW}=\rho_W Vg$.

III. MOTION EQUATION

.Due to the motion equation is nonlinear, it is difficult to obtain analytical solution, this paper adopts the method of numerical calculation.

Simplified equations of motion

Take τ for time step,. when τ tends to zero, within the time of t_k, t_{k+1} ,Force, air density and added mass can be as a constant. According to above, the discussion equation of motion:. Equation can be decomposed into mean equation and the random equation to solve. So the mean equation of motion is:

$$m\ddot{S}+M\ddot{S}=F+G-F_f \quad S(t_k)=S_k \quad \dot{S}(t_k)=\dot{S}_k \quad (2.1)$$

And the random equation is：

$$m\ddot{S}_R+M\ddot{S}_R=F_R \quad S_R(t_k)=0 \quad \dot{S}_R(t_k)=0 \quad (2.2)$$

Solution to mean equation

Within the time of (t_k, t_{k+1}),when $\tau\to 0$,all the forces 、 air density and added mass can be taken as constant， then the original equation can be rewritten as follows [7].

$$\ddot{S}_k=\dot{V}_K=\frac{F+G-F_f}{(m+m_s)} \quad \dot{S}(t_k)=\dot{S}_k=V_k \quad (2.3)$$

Trajectory calculation of aircraft in the air

Taking buoyancy as 0, then.

$$V_{k+1}=V_k+(m+m_s)[|F|_k+|G|_k]\tau$$

Where air density ρ =1.292kg/m^3.

Trajectory calculation of aircraft in the water

If there is no air in the cabin, buoyancy is approximately taken as 0, then

$$V_{k+1}=V_k+(m+m_s)[|F|_k+|G|_k]\tau \quad (2.4)$$

Otherwise,

$$V_{k+1}=V_k+(m+m_s)[|F|_k+|G|_k-|F_f|_k]\tau$$

. Where $F_{fW}=\rho_W Vg$, seawater density is

. $\rho_W=1.025\times10^3 kg/m^3$

following formula to calculate the approximate random displacement are written as :

$$m\ddot{S}_R=F_R \quad S_R(t_k)=0 \quad \dot{S}_R(t_k)=0$$

explicit Euler method to solve the problem, namely

$$V_{k+1}=V_k+\tau\dot{V}_k=V_k+\frac{\tau}{m+m_s}[|F|_k+|G|_k-|F_f|_k]$$

IV. CALCULATION RESULTS

.Taking a common classic small aircraft as an example to determine the expected drop zone by using the general mathematical model, the process is as follows. Hypothesis: the area A of the aircraft is 75m^2. Ay=7m^2, aircraft quality is 2000kg, the initial velocity is (222m/s, 0,0) T ,the

speed of wind is(5,0,0) T and water velocity is 2m/s, the water depth is 600m, the aircraft's lift to drag ratio is 17, according to the above model to calculate the trajectory of aircraft:

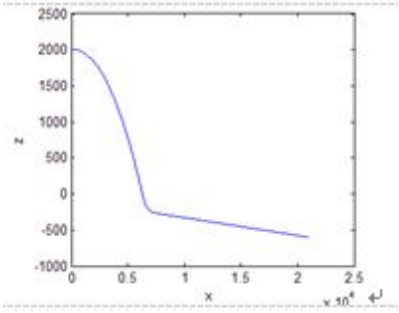


Fig 2. Two-dimensional flight track

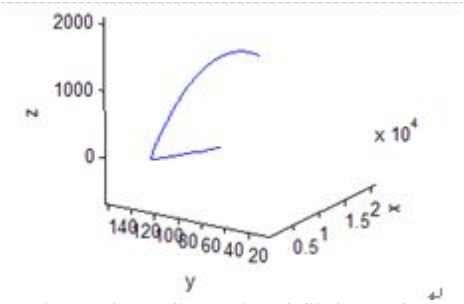


Fig 3. Three-dimensional flight track

Fig .2 is the aircraft trajectory in the XOZ plane, X axis for the longitudinal displacement, Z axis for the height. After losing the power, the plane’ s displacement curve in the air is a parabola, while is a linear drift in the water. The displacement in the water is bigger than in the air. In order to more. accurately describe the way of movement, we give three-dimensional trajectory map as shown Fig 2:

In order to further analyze the trajectory of the aircraft, we extracted velocities in three directions for further study. The aircraft speed is shown in Fig 3.

Fig 3 reflect the change rules of the aircraft speed in the process of falling. vx in the air decreased with height gradually decreased, while vz and vy gradually increased. Falling in to the water, the speed of three directions decreases sharply at the same time, finally the three directional speed is in a stable range.

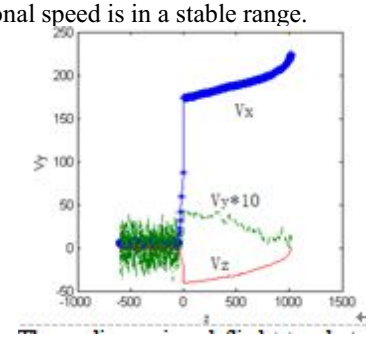


Fig 4. Three-dimensional flight track

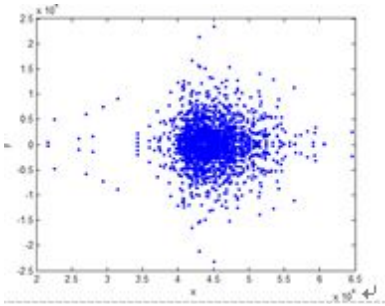


Fig 5. Final landing point distribution of the aircraft

Then we consider factors of wind direction and wind speed, direction and velocity of different currents, with the adjustments of the parameters in the above model, the plane lost distribution map points are as shown in fig4.and the probability distribution in each region of the contour map is shown in fig5.

As shown in Fig 6, final landing point probability distribution contour of aircraft is observed. The shape of final landing point distribution probability is approximately the shape of an ellipse , the longer axis of its ellipse is in the direction of x, namely the flight direction .The red area indicates a high probability of location, blue for low-probability locations. Largest location probability is located in the waters at distance of 33km in front of the loss point, the radius of the region is about 10km, search area 300km ^ 2.

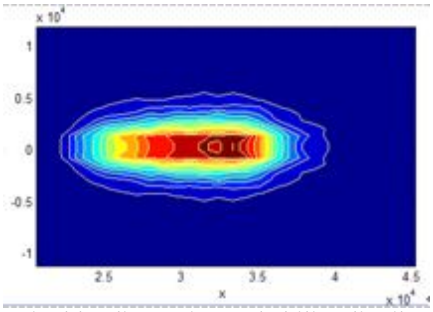


Fig 6. Final landing point probability distribution contour of aircraft

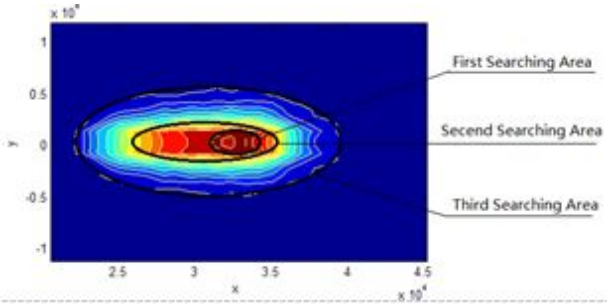


Fig 7. Searching Area

At present, focus should be shifted to search and rescue the aircraft in the shortest time with limited resources. The significance of this distribution probability is that it can help searchers make the right decision the first time to narrow your search range. The regional map is divided into three parts: a high probability distribution area - the first priority

search area; the second highest probability distribution area - the second priority search area; small probability distribution area - close third priority search searching area. If the goal is not found in the first priority search area, search area must be gradually expanded.It is noteworthy that the above conclusion is only the parameters given by the study and can not be a general model to predict aircraft search area. When using this model, based on the specific model of the lost aircraft, and the sea winds, ocean currents and other parameters to determine the specific accurate distribution graphs, so as to provide a more accurate guidance for the search.

.

V. CONCLUSIONS

Based on the principle of fluid mechanics, when the plane lost, the trajectory of the plane are analyzed, and the random equation of motion of the aircraft is given. To make the model more accurate, the model also takes other factors into account such as variation of the air density, added mass of the plane and the buoyancy of water. Then according to the differential thought, the equation of motion is simplified split into mean equation and the stochastic equation. By using Euler method to solve the equation and getting the velocity solution. By integrating the velocity to get the motion trace. Finally we get the trajectory of the aircraft in the air and in water. Through the trajectory equation we can predict the location of the crash which is the prediction model for plane crash area, it is our base model. Finally, we select the model of an aircraft, determine the parameters of the equation and get the regional probability contour map of the lost plane. With the results we get, the search area is divided into three areas named 1 to 3. Partition will provide guidance opinions for the next search.

ACKNOWLEDGMENT

The paper is financially supported by China Ministry of Transport Applied Basic Research Project(No。 2013-329-811-360).

REFERENCES

[1] Le Hardy, Peter K. , Deep ocean search for Malaysia airlines flight 370[J], OCEANS 2014(06), , 2015, 3-7.

[2] Chen, Yean-Ru, Unified security and safety risk assessment - A case study on nuclear power plant[C], Proceedings of 1st International Conference on Trustworthy Systems and Their Applications, TSA 2014, November 13,p 22-28.

[3] Manley, Justin, Advanced technologies for undersea object location[J], OCEANS 2014(06), 2015, 8-12.

[4] Li, Donghui, Propagation regularity of hot topics in Sina Weibo based on SIR model - A simulation research[C], Proceedings - 2014 IEEE Computers, Communications and IT Applications Conference, ComComAp 2014, p 310-315

[5] Su, Yuanchao, A target detection method with morphological knowledge for high-spatial resolution remote sensing image applying for search and rescue in aviation disaster[C], Proceedings of SPIE - The International Society for Optical Engineering, v 9260, 2014.

[6] Xin Chen, Mengyu Li, and Shesheng Zhang，A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil［J］, Journal of Algorithms & Computational Technology,2014,Vol. 8 No. 3, 249-266.

[7] Xin Chen, Shengping Jin, Shesheng Zhang，Dan Li，A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil［J］, Journal of Algorithms & Computational Technology,2015,Vol. 9 No. 2, 163-174. .

Simple Computational Methods for Large Deformation of Plate-spring End Imposed by Varying Load

Jun Zhang[1,2]
1.Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment & Technology
2.School of Mechanical Engineering, Jiangnan University, Wuxi, P.R.China
jzhang@jiangnan.edu.cn

Guangyuan Liu[1,2]
1.Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment & Technology
2.School of Mechanical Engineering, Jiangnan University, Wuxi, P.R.China
hap-cleoy@163.com

***Abstract*—In this paper, a pneumatic flexible finger joint is recommended, which uses the plate-spring as skeleton. Bending status of the plate-spring is equal to the large deflection of elastic cantilever, which is actuated by a pneumatic artificial muscle actuator of longitudinal expandable rubber bellows. By this actuator, a follower force tangential to the axis of the plate-spring as well as a moment is imposed at the end of the beam. The equations to define the relation between deformation and air pressure in the actuator have been established, and mathematical methods have been derived. Two numerical solutions have been given severally by means of software Mathematica5 and Matlab, and calculation results show the error between two methods is tiny.**

Keywords-Elastic Cantilever; Large Deformation; Varying Load; Elliptic Integral method; differential method

I. INTRODUCTION

Service robot, bionic robot and agricultural robot must contact with organisms directly, so they have to require high security and adaptability. A pneumatic flexible finger joint is described now. The plate-spring possesses a flexible skeleton. Its bending status is equal to the large deformation of a elastic beam. The bend angle of the joint is related to air pressure in compliant artificial muscle actuator, a longitudinal expandable rubber bellows.

The study of large deformation has mainly focused on numerical analysis. Davoodinik[1] with some scholars has systematically introduced the method which solved the problem of large deformation with nonlinear finite element. Based on measurement of the Jaumann stress and strain, accurate coordinate conversion, and a new theory of the virtual orthogonal rotation, PAI etc[2] derived a complete Lagrangian finite element scheme to solve the problem of large deformation beam. ZUPAN etc[3] proposed a finite element method based on the theory of accurate geometry finite strain beam. Another major investigation is about compliant mechanism from the Euler–Bernoulli beam theory taking the geometric non-linearity into account. Solution to the resulting non-linear differential equation has been obtained in terms of elliptic integrals of the first and second kind[4]. The large deflection of a cantilever beam under a combined loading was investigated and the numerical solution was obtained by using Butcher's fifth order Runge–Kutta method[5]. Numerical schemes have also been proposed where the forces along with moments are applied only at the free end[6]. The occurrence of any inflection point within the beam segment requires special attention. A model for a cantilever beam with end moment acting in the opposite direction as the end force is given, which may or may not cause an inflection point[7]. The elliptic integral solutions are used to determine when an inflection point will exist. The beam end deflections are then parameterized using a different parameterization from previous models, which renders the deflection paths easier to model with a single degree of freedom system. Forces and moments also exist at some intermediate locations. More recently, two simple methods[8] have been proposed to obtain large deflection of a cantilever beam including geometric non-linearity, one numerical method is called non-linear shooting and another semi-analytical method is known as Adomian decomposition.

The aforementioned articles dealt only with unchanged dimension and direction. In this article, the dimension and direction of the force and even the dimension of the moment made by the force have all been varying with the air pressure in artificial muscle actuator, and two simple and direct solutions are given. Scientific computation software, Mathematica5 and Matlab, are used. Results of two solutions are analysed and commented.

II. STRUCTURE AND STATIC MODEL OF JOINT WITH PLATE-SPRING SKELETON

The joint structure is shown in Fig.1. The rubber bellows 6 (longitudinal section as letter V), is clamped on the sinuate mouth of caput seat 1 and tail seat 7 by holder 4. Both ends of plate-spring 5 are fixed on caput seat 1 and tail seat 7 by pads. A sealed cavity is formed with bellows 6, caput seat 1 and tail seat 7, as a compliant pneumatic artificial muscle actuator [9]. Gas flows into the sealed cavity through pipe coupling 9 and pipe 10. The pipe coupling 9 is clamped on tail seat 7. The flexible annular woof fiber 11 inside of rubber bellows 6 are entwined with flexible longitudinal fibers 12 distributing along its longitudinal profile, and fiber grid with enhancement are made up. The thickness of backup plate 2 can be adjusted according to different expects, so that eccentricity e and bend moment are changed.

As shown in Fig.1, the radial or longitudinal dimensions of parts are equal and connected to both sides of the plate-

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.107

spring. Reasonable assumptions can be made based on the structure of the joint: (1) The force on caput seat and tail seat from elastic bend of the bellows can be negligible because elastic bend of the bellows is much easier. (2) Under the conditions of the force, the length of the plate spring is equal to the original length L. (3) All parts except plate-spring and the bellows are assumed to be rigid. (4) Weight and inertia of all parts are negligible.

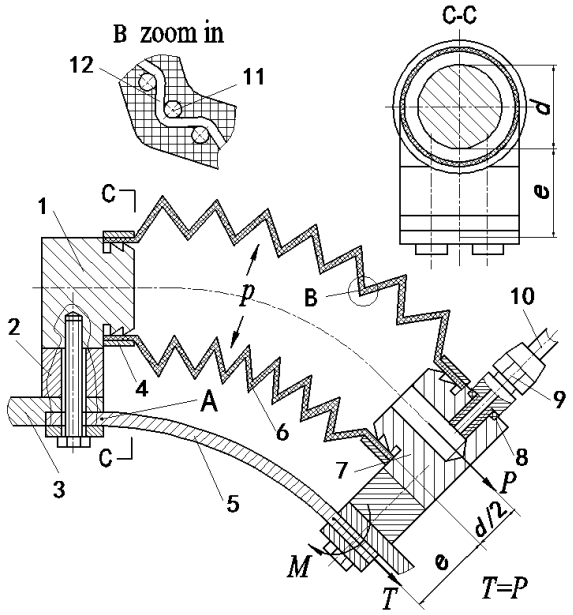


Figure 1. Structure of the flexible joint

As shown in Fig.1, the radial or longitudinal dimensions of parts are equal and connected to both sides of the plate-spring. Reasonable assumptions can be made based on the structure of the joint: (1) The force on caput seat and tail seat from elastic bend of the bellows can be negligible because elastic bend of the bellows is much easier. (2) Under the conditions of the force, the length of the plate spring is equal to the original length L. (3) All parts except plate-spring and the bellows are assumed to be rigid. (4) Weight and inertia of all parts are negligible.

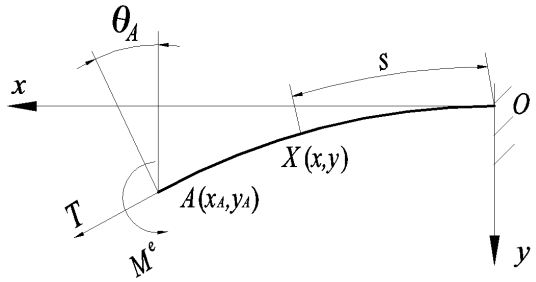


Figure 2. Force and deformation of finger plate-spring

Force and deformation of finger plate-spring are given in Fig.2. The plate-spring can be simplified as cantilever AO. Taking no account of the elastic stiffness of rubber bellows, the plate-spring will be the only flexible element, and its large deformation is clearly equal to the large deformation of the cantilever. The force *P*, acted at the center of caput seat, is $P=\pi(d/2)^2 p$, if there has a pressure *p* in rubber bellows. The force P makes the cantilever AO bend.

At the initial point A of plate-spring bend (end of cantilever AO), the force *T* (along the tangent of point A) and moment M^e are driven by force *P*. Where: $T=\pi(\frac{d}{2})^2 p$, $M^e=\frac{\pi}{4}d^2(e+\frac{d}{2})p$. *d* is the diameter of wave hollow circle of the bellows, *e* is eccentricity, *p* represents air pressure, as shown in Fig. 1.

Bend of the flexible finger joint can be translated into the solution of bend deformation of plate-spring. Actually it is the solution to the coordinate x_A, y_A , and slope θ_{A} (viz. bend angle) of point A after large deflection of cantilever. At a random point X of cantilever, the bend moment *M* is

$$M=M^e+T\sin\theta_{\mathrm{A}}(x_A-x)-T\cos\theta_{\mathrm{A}}(y_A-y) \quad (1)$$

where: $b=1/EI$, EI is bend rigidity. At a random point X (where: *s* is the length of arc XO) of the cantilever, Using the Euler–Bernoulli moment–curvature relationship, we can conclude that

$$\frac{\mathrm{d}\theta}{\mathrm{d}s}=bM(s)=b[M^e+T\sin\theta_{\mathrm{A}}(x_A-x)-T\cos\theta_{\mathrm{A}}(y_A-y)] \quad (2)$$

After differentiating Eq. (2) by *ds*, and substituting $\frac{dx}{ds}=\cos\theta$, $\frac{dy}{ds}=\sin\theta$. Eq. (3) can be derived as

$$\frac{d^2\theta}{ds^2}=bT(\sin\theta_A\cos\theta-\cos\theta_A\sin\theta)=bT\sin(\theta_A-\theta) \quad (3)$$

III. The Method of Elliptic Integral

By multiplying $d\theta$ to Eq.(3), we may have $d(\frac{d\theta}{ds})\frac{d\theta}{ds}=bT\sin(\theta_A-\theta)d\theta$. After its integrating, we also have $\frac{1}{2}(\frac{d\theta}{ds})^2=bT\cos(\theta_A-\theta)+C$, C is constant parameter, from Eq. (2), the following boundary condition is $\frac{d\theta}{ds}\Big|_{s=L}=bM^e$. So we have $C=\frac{1}{2}(bM^e)^2-bT$, and equation (4) can be obtained

$$\frac{d\theta}{ds}=\sqrt{(bM^e)^2-2bT+2bT\cos(\theta_A-\theta)} \quad (4)$$

According to the assumption: the length *L* of the plate-spring does not change, Eq. (5), (6), (7) can be derived as

$$L=\int_0^L ds=\int_0^{\theta_A}\frac{1}{\sqrt{(bM^e)^2-2bT+2bT\cos(\theta_A-\theta)}}d\theta \quad (5)$$

$$x_A=\int_0^{x_A}dx=\int_0^L\cos\theta\cdot ds=\int_0^{\theta_A}\frac{\cos\theta}{\sqrt{(bM^e)^2-2bT+2bT\cos(\theta_A-\theta)}}d\theta \quad (6)$$

$$y_A=\int_0^{y_A}dy=\int_0^L\sin\theta\cdot ds=\int_0^{\theta_A}\frac{\sin\theta}{\sqrt{(bM^e)^2-2bT+2bT\cos(\theta_A-\theta)}}d\theta \quad (7)$$

The elements T, M^e and b will be known, when the pressure p and plate-spring are determined. The calculations about manipulator's deformation refer to three elements. They are θ_A, x_A, y_A, and limited by equations (5), (6), (7). Because equations (5), (6), (7) are nonlinear integration and the elements can not be uniquely worked out in a straightforward way due to θ_A which is in the integrated expression and upper integration range value. The elliptic integral method is given in detail by reference [10].

A. Angle Transformation Equation of Slope

Using equation of trigonometry $\cos 2\alpha = 1 - 2\sin^2\alpha$, Equation (5) can be re-written as

$$L = \int_0^{\theta_A} \frac{1}{\sqrt{(bM^e)^2 - 4bT\sin^2(\frac{\theta_A - \theta}{2})}} d\theta \tag{5a}$$

Let $\phi = \frac{\theta_A - \theta}{2}$, and substitute the elements T, M^e and b into Eq. (5a), Eq. (5b) can be obtained as

$$L = -c \cdot \int_{\frac{\theta_A}{2}}^{0} \frac{1}{\sqrt{1 - k^2\sin^2\phi}} d\phi = c \cdot \int_0^{\frac{\theta_A}{2}} \frac{1}{\sqrt{1 - k^2\sin^2\phi}} d\phi \tag{5b}$$

Where: $c = \frac{16EI}{\pi d^2(2e+d)p}$, $k^2 = \frac{64EI}{\pi d^2(2e+d)^2 p}$,

Let $k_1 = \frac{1}{k}$, (where $k_1^2 < 1$), $\sin\beta = k\sin\phi$, we have, Equation (5b) can be modified to standard form of elliptic integral of first type

$$L = c \cdot \int_0^{\beta_A} \frac{k_1}{\cos\phi} d\beta = ck_1 \cdot \int_0^{\beta_A} \frac{1}{\sqrt{1 - k_1^2\sin^2\beta}} d\beta = ck_1 \cdot F(k_1, \beta_A) \tag{5c}$$

Where $E(k_1, \beta_A)$ is Legendre elliptic integral of first type [10], and its integration range β_A can be obtained from $k_1 \sin\beta_A = \sin(\theta_A/2)$. For a certain air pressure p, with structure dimensions being determined, assuming θ_A, we can find β_A, c, k_1, $F(k_1, \beta_A)$. Fortunately, using Step-by-Step Approximation method, numerical value of slope θ_A can be obtained from Elliptic F function of software Mathematica, when the precision is sufficient according to Eq. (5c). In Mathematica software, the machine precision is 16 digits. Although its standard output displays only the first 6 digits, it is enough to meet the application requirements.

B. Calculation of Displacement Coordinate

Let $\phi = \frac{\theta_A - \theta}{2}$, and substitute the elements T, M^e and b into Eq. (6) and Eq. (7), Eq. (6a) and Eq. (7a) can be found as

$$x_A = c \cdot \int_0^{\frac{\theta_A}{2}} \frac{\cos(\theta_A - 2\phi)}{\sqrt{1 - k^2\sin^2\phi}} d\phi = c \cdot \int_0^{\frac{\theta_A}{2}} \frac{\cos\theta_A\cos 2\phi + \sin\theta_A \sin 2\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi \tag{6a}$$

$$y_A = c \cdot \int_0^{\frac{\theta_A}{2}} \frac{\sin(\theta_A - 2\phi)}{\sqrt{1 - k^2\sin^2\phi}} d\phi = c \cdot \int_0^{\frac{\theta_A}{2}} \frac{\sin\theta_A\cos 2\phi - \cos\theta_A \sin 2\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi \tag{7b}$$

Let $\sin\phi = x$, $k_1 = \frac{1}{k}$, $\sin\beta = k\sin\phi$, $d\phi = k_1 \frac{\cos\beta}{\cos\phi} d\beta$, we have

$$\int_0^{\frac{\theta_A}{2}} \frac{\sin 2\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi = \int_0^{\frac{\theta_A}{2}} \frac{2\sin\phi\cos\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi =$$

$$2k_1 \int_0^{\sin\frac{\theta_A}{2}} \frac{x}{\sqrt{k_1^2 - x^2}} dx = 2k_1 (-\sqrt{k_1^2 - x^2}) \Big|_0^{\sin\frac{\theta_A}{2}} = 2k_1 (k_1 - \sqrt{k_1^2 - \sin^2\frac{\theta_A}{2}}) \tag{8}$$

$$\int_0^{\frac{\theta_A}{2}} \frac{\cos 2\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi = \int_0^{\frac{\theta_A}{2}} \frac{1 - 2\sin^2\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi$$

$$= \int_0^{\frac{\theta_A}{2}} \frac{1}{\sqrt{1 - k^2\sin^2\phi}} d\phi + \frac{2}{k^2}[\int_0^{\frac{\theta_A}{2}} (\sqrt{1 - k^2\sin^2\phi} - \frac{1}{\sqrt{1 - k^2\sin^2\phi}}) d\phi]$$

$$= (1 - \frac{2}{k^2}) \int_0^{\frac{\theta_A}{2}} \frac{1}{\sqrt{1 - k^2\sin^2\phi}} d\phi + \frac{2}{k^2} \int_0^{\frac{\theta_A}{2}} \sqrt{1 - k^2\sin^2\phi} d\phi$$

$$= (1 - \frac{2}{k^2}) \frac{L}{c} + \frac{2}{k^2} \int_0^{\frac{\theta_A}{2}} \sqrt{1 - k^2\sin^2\phi} d\phi \tag{9}$$

Similarly, Equation (9) can be modified to standard form of elliptic integral of second type

$$\int_0^{\frac{\theta_A}{2}} \frac{\cos 2\phi}{\sqrt{1 - k^2\sin^2\phi}} d\phi = (1 - 2k_1^2) \frac{L}{c} + 2k_1^2 \int_0^{\beta_A} \frac{k_1(1 - \sin^2\beta)}{\sqrt{1 - k_1^2\sin^2\beta}} d\beta$$

$$= (1 - 2k_1^2) \frac{L}{c} + 2k_1^3 \int_0^{\beta_A} \frac{1}{\sqrt{1 - k_1^2\sin^2\beta}} d\beta - 2k_1^3 \int_0^{\beta_A} \frac{\sin^2\beta}{\sqrt{1 - k_1^2\sin^2\beta}} d\beta$$

$$= (1 - 2k_1^2) \frac{L}{c} + 2k_1^3 \cdot \frac{L}{ck_1} + 2k_1^3 \frac{1}{k_1^2} [\int_0^{\beta_A} (\sqrt{1 - k_1^2\sin^2\beta} - \frac{1}{\sqrt{1 - k_1^2\sin^2\beta}}) d\beta]$$

$$= \frac{L}{c} - 2k_1 \cdot \frac{L}{ck_1} + 2k_1 \int_0^{\beta_A} \sqrt{1 - k_1^2\sin^2\beta} d\beta = -\frac{L}{c} + 2k_1 \cdot E(k_1, \beta_A) \tag{10}$$

where: $k_1 \sin\beta_A = \sin\frac{\theta_A}{2}$, $E(k_1, \beta_A)$ is Legendre elliptic integral of second type [10]. From Eqs. (6a), (7a), (8), (9), (10), the coordinate of Eq.6b and Eq.7b can be found as

$$x_A = c\cos\theta_A[-\frac{L}{c} + 2k_1 \cdot E(k_1, \beta_A)] + c\sin\theta_A \cdot 2k_1(k_1 - \sqrt{k_1^2 - \sin^2\frac{\theta_A}{2}}) \tag{6b}$$

$$y_A = c\sin\theta_A[-\frac{L}{c} + 2k_1 \cdot E(k_1, \beta_A)] - c\cos\theta_A \cdot 2k_1(k_1 - \sqrt{k_1^2 - \sin^2\frac{\theta_A}{2}}) \tag{7b}$$

For certain air pressure p, with the parameter θ_A obtained from Eq. (5c) and known elements β_A, c, k_1, $E(k_1, \beta_A)$ is obtained by means of Elliptic E function of software Mathematica. The coordinates x_A and y_A can be obtained from Eqs. (6b) and (7b).

IV. THE METHOD OF DIFFERENTIAL

A. The Solution of Angle Equation

Eq. (3) is a nonlinear differential equation of the boundary value problem (BVP), and can be given as Eq. (11)

$$\text{D.E}\quad \frac{d^2\theta}{ds^2} = bT\sin(\theta_A - \theta) \tag{11}$$

$$\text{B.C}\ \ \theta|_{s=0} = 0,\ \theta|_{s=L} = \theta_A,\ \frac{d\theta}{ds}\Big|_{s=L} = bM^e$$

Because Eq. (11) cannot be uniquely concluded in a straightforward way due to θ_A which is in the differential expression and its upper boundary-value. According to one proper answer in mathematical solution of engineering, we suppose: Eq. (11) has only one solution. For a certain pressure p, assuming U, substitute for parameter θ_A, Eq. (11) is converted into $\frac{d^2\theta}{ds^2} = bT\sin(U-\theta)$. Work out the $\tilde{\theta}_A$ on the upper boundary ($s = L$) of equation $\frac{d^2\theta}{ds^2} = bT\sin(U-\theta)$ by means of bvp4c() function of software Matlab. Compare $\tilde{\theta}_A$ with U, and then adjust U using method of step-by-step approximation. When the precision is $\left|\tilde{\theta}_A - U\right| \le 10^{-6}$, we consider that $\tilde{\theta}_A$ is equal to θ_A.

B. The Solution of Displacement Coordinates Equation

$\frac{d\theta}{ds} = \cos\theta\frac{d\theta}{dx} = \sqrt{(bM^e)^2 - 2bT + 2bT\cos(\theta_A - \theta)}$, and $\frac{d\theta}{dx}\Big|_{x=0} = \sqrt{(bM^e)^2 - 2bT + 2bT\cos\theta_A}$ can be had from Eq. (4).

$$\frac{d}{ds}(\frac{d\theta}{ds}) = \cos\theta\frac{d}{dx}(\cos\theta\frac{d\theta}{dx}) = \cos\theta[-\sin\theta(\frac{d\theta}{dx})^2 + \cos\theta\frac{d^2\theta}{dx^2}]$$
$= bT\sin(\theta_A - \theta)$ from Eq.(3), we get

$$\text{D.E}\quad \frac{d^2\theta}{dx^2} - \tan\theta(\frac{d\theta}{dx})^2 - \frac{bT\sin(\theta_A - \theta)}{(\cos\theta)^2} = 0 \tag{12}$$

$$\text{I.C}\quad \theta|_{x=0} = 0,\ \frac{d\theta}{dx}\Big|_{x=0} = \sqrt{(bM^e)^2 - 2bT + 2bT\cos\theta_A}$$

Equation (12) is a differential equation of an initial value problem (IVP) after has been obtained, and can be worked out by means of ode45 function of software Matlab.

When $y = 0$, $\theta = 0$, from Eq. (4), we have $\frac{d\theta}{dy}\Big|_{y=0} = \frac{\sqrt{(bM^e)^2 - 2bT + 2bT\cos\theta_A}}{\sin(10^{-10})}$ (Because the denominator of the factor can not be 0 in numerical calculation, we set $\theta = 10^{-10}$).

In the same way, from Eq. (3), we have: $\frac{d}{ds}(\frac{d\theta}{ds}) = \sin\theta\frac{d}{dy}(\sin\theta\frac{d\theta}{dy}) = \sin\theta[\cos\theta(\frac{d\theta}{dy})^2 + \sin\theta\frac{d^2\theta}{dy^2}]$
$= bT\sin(\theta_A - \theta)$, and we can have

$$\text{D.E}\quad \frac{d^2\theta}{dy^2} + \cot\theta(\frac{d\theta}{dy})^2 - \frac{bT\sin(\theta_A - \theta)}{(\sin\theta)^2} = 0 \tag{13}$$

$$\text{I.C}\quad \theta|_{y=0} = 0,\ \frac{d\theta}{dy}\Big|_{y=0} = \frac{\sqrt{(bM^e)^2 - 2bT + 2bT\cos\theta_A}}{\sin(10^{-10})}$$

Equation (13) is a differential equation of an initial value problem (IVP) after θ_A has been obtained, and y_A can also be worked out by means of ode45 function of software Matlab.

V. RESULTS

Assume structure dimensions $L = 0.016$ m, $d = 0.015$ m, $e = 0.006$ m, and the rectangle section plate-spring with modulus of elasticity $E = 19.65\times10^7$ N/m2 and with breadth $B = 0.009$ m, thickness $h = 0.0005$ m.

By two computational methods, the slope θ_A (viz. bend angle of joint), coordinates x_A and y_A of cantilever end with different air pressure are given in Table 1.

By the two differential methods we can get the deformation status (four parameters θ, x, y and s, when p is given) of an random point of the plate-spring, and the relation curves of $s\sim\theta$, $s\sim x$, $s\sim y$, $x\sim\theta$, $y\sim\theta$, $x\sim y$ also can be given. Relation curves between slope and length s of the plate-spring are given in Fig. 3.

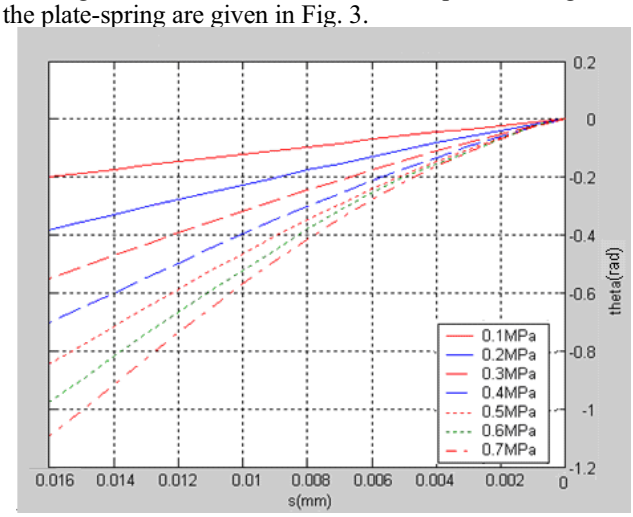


Figure 3. Slope of random point in plate-spring with different air pressure

VI. CONCLUSIONS

Further work to be done is to investigate the choice of rubber and fiber materials, strength calculation of fabric reinforced composite material (relation with load-bearing pressure of the bellows), technology of manufacture, and experimental study on applied status with skeleton of plate-spring.

Dimension and direction of the force, and dimension of the moment continuously vary with pneumatic pressure in cavity of the actuator imposed at the end of the cantilever, whose state has never been found in the past. Two methods have been created in the study. Mathematica5 and Matlab are severally used in the calculation method.

The calculation results mentioned above show that the error between the two methods is extremely tiny and fully prove the correctness of the two kind of calculation methods.

We can only find the status (θ_A, x_A, y_A) of cantilever end by the elliptic integral method which also needs the help of Matlab software to calculate differential equations .However, by the differential method not only the status and trace of motion of the cantilever end but also the status at a random point of plate-spring can be found. Moreover by contrast, the differential method is more concise and makes clearer analysis of elastic cantilever at any point.

Besides, Conclusion is that the differential method is a more simple numerical calculation method of elastic cantilever, which has big significance on the study of large deflection deformation.

REFERENCES

[1] Belytschko.T, Liu.W.K and Moran.B, "Nonlinear Finite Elements for Continua and Structures", 2000. Chichester, John Wiley and Sons , pp. 356–370.

[2] Pai.P.F, Anderson.T and J.Wheater, "Large-Deformation Tests and Total-Lagrangian Finite-Element Analyses of Flexible Beams", J International Journal of Solids and Structures, 2000, vol.37, pp2951-2980.

[3] Zupan.D and Saije.M, "Finite-Element Formulation of Geometrically Exact Three- Dimensional Beam Theory Based on Interpolation of Strain Measures", J Computer Methods in Applied Mechanics and Engineering, 2003, vol.192, pp5209-5248

[4] A.R.Davoodinik and G.H.Rahimir, "Large deflection of flexible tapered functionally graded beam", Acta Mechanica Sinica, 2011, vol.5, pp767-777.

[5] Kyungwoo Lee, "Large defl¬ections of cantilever beams of non-linear elastic material under a combined loading", International Journal of Non-Linear Mechanics, 2002, vol.37, pp439–443

[6] A. Saxena and S.N. Kramer, "A simple and accurate method for determining large deflections in compliant mechanisms subjected to end forces and moments", ASME J. Mech. Des, 1998, vol.120, pp392–400

[7] C. Kimball and L.W. Tsai, "Modeling of flexural beams subjected to arbitrary end loads", ASME J. Mech. Des, 2002, vol.124, pp223–234

[8] A. Banerjee, B. Bhattacharya, A.K. Mallik; 'Large deflection of cantilever beams with geometric non-linearity: Analytical and numerical approaches', International Journal of Non-Linear Mechanics, 2008, vol.43, pp366-376.

[9] J.Zhang and Q.J.Zhang, "Simulating of Muscle Joint Directly or Indirectly Driven by Pneumatic Pressure", Acta Mechanica, 2008, vol.197, pp119-130

[10] Jun.Zhang and Qiu J.Zhang, "BENDING OF FLEXIBLE JOINT DRIVEN BY LINEAR EXPANDABLE ARTIFICIAL MUSCLE", International Journal of Modelling and Simulation, 2011, vol 31, pp1-5.

TABLE I. BEND STATUS OF THE END OF ELASTIC CANTILEVER (PLATE-SPRING) WITH DIFFERENT AIR PRESSURE

p (MPa)		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
θ_A	**1**	0	11.3928	21.8617	31.4697	40.2902	48.4025	55.8883	62.8293
(°)	**2**	0	11.3928	21.8617	31.4697	40.2902	48.4025	55.8883	62.8293
x_A	**1**	16	15.8977	15.6370	15.2765	14.8621	14.4263	13.9909	13.5688
(mm)	**2**	16	15.9180	15.6378	15.2768	14.8625	14.4267	13.9916	13.5696
y_A	**1**	0	1.5524	2.8881	4.0109	4.9372	5.6887	6.2902	6.7657
(mm)	**2**	0	1.5524	2.8881	4.0110	4.9373	5.6887	6.2908	6.7664

Attention: No1 is the elliptic integral method; No2 is the differential method.

Large ship fluid-structure coupling deformation calculation based on large deviation theory

Xinyun Liu
School of Mathematics and Statistics, Wuhan University,430070,Chinan
e-mail: LiuXinyun@qq.com

XinCong Zhou, 2Shesheng Zhang Yuguang Li
Wuhan University of Technology, Wuhan, 430070
e-mail: Liyuguang@qq.com

***Abstract*—study of ship deformation accumulation for a long time is of practical significance. According to the initial-boundary value problem of the deformation of the ship, large deviation theory gives the deformation time series convergence theorem, which deduces the convergence rate formula of discrete deformation. The convergence speed of the actual example is attached**

Keywords-Ship, deformation, large deviation, the time series

I. INTRODUCTION

.In a rough environment, large ship sailing on the oceans will occur deformation. The distortion movement for a long period of time, will produce a metal fatigue effect. The effects of fatigue, not only reduce the sailing speed, but also lead to mechanical failure of the ship. Machine failure is out of control is one of the main hazards for water carriage safety transportation. Thus the research of on ship deformation is necessary [1].

Scholars in home and aboard developed a class of orthogonal cutting with radial basis function (RBF) grid method and mathematical models that present a type of effective and high precision nonlinear calculation method to solve ship and ocean engineering complex multi-body mobile problems[2]; through modeling and numerical simulation of wave motion of the large oil tanker, calculating the resonance response of large oil tanker in the regular wave succeeded [3]. [4] studied coupling power transfer mechanism and reveals the coupling mechanism of the hull deformation from four aspects: the green's function's intrinsic characteristics and quick calculation method, theory of ship section and fluid-solid interaction in time and space, free surface of the hull fluid-solid interaction dynamics variational principle, and the large ship propulsion system and the dynamic coupling hull VOSS mapping theory.When the ship deformation time series is gotten, we can do convergence analysis by the large deviation theory [5]. Large deviation theory as a refinement of the large Numbers theorem is applied to the convergence problem different from the central limit theorem, which gets fast development in recent years[6].

By using large deviation theory, this paper will consider deformation of large ships moving on the wave water. Section II will give basic theory of ship’ s large deviation; Section III will give initial boundary problem; section IV will calculate conversion speed of time series.

II. LARGE DEVIATION THEORY ON SHIP DEFORMATION

.Because the wave forces to the ship are random, the vessel deformation under the action of wave force Y(t) is random too. At time t_k, the measurement of the deformation of ship is $Y_k=Y(t_k)$. Assume that Y_k are mutually independent and identically distributed random variables series, whose distribution is P.We take $y_n=\frac{1}{n}\sum_{k=1}^{n}Y_k$ with distribution. P_k The following research is the convergence speed of the series $\{P_k\}$.

Usually, $\{u_k\}$ is gens probability measure. If the function I(*):X→$[0,\infty]$meets the conditions:

(1) I(*) is not identical to zero, lower semicontinuous, convex function;

(2) For any constant l , {x:I(x)<l} is set compact tin X;

(3) Any closed set F of X,

$$\overline{\lim_{\varepsilon->0}}\ \varepsilon\log u_n(F)\le-\inf_{x\in F}I(x)$$

(4) Any open set G of X,

$$\underline{\lim_{\varepsilon->0}}\ \varepsilon\log u_n(G)\ge-\inf_{x\in G}I(x)$$

Say that $\{u_n,n>0\}$ meets the large deviation principle, with rate function I(*). When the upper and lower limits are equal, then I(*) is the convergence rate function.

Theorem 1, Assuming that ship deformation is independent, identically distributed random variables series, with distribution P ,taking $y_n=\frac{1}{n}\sum_{k=1}^{n}Y_k$ with distribution. P_n If generating function $M(\theta)=E(e^{\theta x_k})$ is limited, $\{P_n\}$satisfies large deviation principle, with rate function

$$I(x)=\sup_{\theta\in R}[\theta x-\log M(\theta)]\qquad x\in R \tag{2.1}$$

With the above theorem, we have the following corollaries.

Corollary 1: If the ship deformation is independent, identically distributed random variables series, with the distribution P, and satisfies the large deviation principle.

Pf: in the large deviation principle, when n is large enough, take ε =1/n. .Because the rate function exists, upper and lower limits exist. Thus, inference is established. And the proof is complete.

Corollary 2: If the ship deformationis standard normal distribution N(0,1), and independent with each other, the rate of the function is:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.108

$$I(x)=\frac{x^2}{2} \tag{2.2}$$

Pf: Due to ship deformation obeys the standard normal distribution . It is independent and identically distributed random variables series, with the distribution of p. The generating function of standard normal distribution function is $M(\theta)=E(e^{\theta x_k})=e^{\frac{\theta^2}{2}}$:,

so.

$$I(x)=\sup_{\theta\in R}[\theta x-\log M(\theta)]=\sup_{\theta\in R}[\theta x-\frac{\theta^2}{2}]=\frac{x^2}{2}$$

Corollary 3, If the ship deformation obeys the standard normal distribution $N(\mu,\sigma^2)$ and are independent of each other, the rate function:

$$I(x)=\frac{(x-\mu)^2}{2\sigma^2} \tag{2.3}$$

Pf: the ship deformation obeys the standard normal distribution and is independent, identically distributed random variables series, with distribution of p. The generating function of normal distribution is $M(\theta)=e^{\mu\theta+\frac{\sigma^2\theta^2}{2}}$, so .

$$I(x)=\sup_{\theta\in R}[\theta x-\log M(\theta)]=\sup_{\theta\in R}[\theta x-\mu\theta-\frac{\sigma^2\theta^2}{2}]$$

We set $G=x\theta-\mu\theta-\frac{\sigma^2\theta^2}{2}$. Make its derivative is zero ,so

$$G'=x-\mu-\sigma^2\theta=0$$

, thus $\theta=\frac{x-\mu}{\sigma^2}$, put it into the original equation, to get:

$$I(x)=\frac{x-\mu}{\sigma^2}[x-\mu]-\frac{\sigma^2}{2}(\frac{x-\mu}{\sigma^2})^2=\frac{(x-\mu)^2}{2\sigma^2}$$

$$I(x)=\frac{(x-\mu)^2}{2\sigma^2}$$

Corollary 4: If the ship deformationhas discrete distribution, independent of each other,in the closed set F, the rate function is:

$$I(x)=\max_{\theta\in F}[\theta x-\log(\frac{1}{N}\sum_{i=1}^{N}e^{\theta y_i}p_i)]\qquad x\in R$$

Pf: The ship deformation is independent and has discrete distribution $p_i=P\{y=y_i\}$, So the generating function is $M(\theta)=E(e^{\theta x_k})=\frac{1}{N}\sum_{i=1}^{N}e^{\theta y_i}p_i$ Substitute it to rate function, and consider in the closed set F, there is

$$.\ I(x)=\max_{\theta\in F}[\theta x-\log(\frac{1}{N}\sum_{i=1}^{N}e^{\theta y_i}p_i)]\qquad x\in R \tag{2.4}$$

III. THE INITIAL-BOUNDARY VALUE PROBLEM OF LARGE SHIP DEFORMATION

.A large ship sails on the sea waves, under the effect of fluid and solid coupling, whose initial-boundary value problem is:

$$\begin{cases}\dfrac{\partial^2 w}{\partial t^2}-c^2\dfrac{\partial^2 w}{\partial x^2}+bw=f(x,t)\\ w(x,t=0)=0\qquad \dfrac{\partial w(x,t=0)}{\partial t}=0\\ \dfrac{\partial w(x=0,t)}{\partial x}=0\qquad \dfrac{\partial w(x=L,t)}{\partial t}=0\end{cases} \tag{3.1}$$

Where t is time variable. x is the variable in distance from cross section and w=w(x,t) is the ship deformation function. c,b are constants. L is ship’s length. F(x,t) is a fluid-structure coupling reaction.The solution of equation is

$$w=\sum_{k=0}^{\infty}S_k^*(t)[1-\cos\frac{k\pi}{L}x]\frac{L}{k\pi}+g(t) \tag{3.2}$$

Where

$$S_k^*(t)=\frac{2/L}{q_k}\int_0^t\int_0^L f_y(y,\tau)\sin\frac{k\pi y}{L}dy\sin q_k(t-\tau)d\tau$$

$$q_k=\sqrt{b+\frac{c^2k^2\pi^2}{L^2}}$$

Here f_y is derivative of function of y. The initial value of function can be gotten from the following equations.

$$f(0,t)+g''(t)-c^2v_x(0,t)+bg(t)=0$$

$$g(0)=0\qquad g'(0)=0 \tag{3.3}$$

the solution to the initial value problem of ship deformation is

$$u=\sum_{k=0}^{\infty}u_k[1-\cos\frac{k\pi}{L}x]\frac{L}{k\pi}+g(t)$$

$$u_k=q_{k1}C(q_kt)-q_{k1}C(\omega t)+q_{k2}S(q_kt)-q_{k3}S(\omega t)$$

Where

$$g=-c^2\sum_{k=0}^{\infty}g_k\frac{k\pi}{L}-G_c+H_1S_g(\omega t)$$

$$g_k=q_{k1}C_g(q_kt)-q_{k1}C_g(\omega t)+q_{k2}S_g(q_kt)-q_{k3}S_g(\omega t)$$

.

IV. LARGE DEVIATION CALCULATION

.Assume that the time series of wave height measured in a fixed point is $\{z_k\}$. We can get wave ' s spectrum $S = S(\omega)$ by Fourier transform from $\{z_k\}$. Then amplitude can be calculated by $H = [2S(\omega)\Delta\omega]^{1/2}$, where ω is the step size of circular frequency . Wavelength λ and wave number P can be calculated by

$$\lambda = \frac{2\pi g}{\omega^2} \qquad K = \frac{2\pi}{\lambda} \tag{4.1}$$

Set $\omega = \omega_j, j = 1,2,...$, then the sea wave can be expressed in a form of series: $z = \sum_{j=1}^{\infty} H_j \sin(K_j x - \omega_j t)$. At time t_k, the ship deformation caused by sinusoidal wave with circular frequency ω_k is w_k . $\{w_k\}$ is a time series. At the initial time t=0, choose N circular frequency values. Calculate the ship deformation value at a fixed time . We can get the probability distribution of ship deformation according to statistical theories. Then the generating function can be calculated by , where is the number of measured values. Then compute the maximum :

$$I(x.\theta) = [\theta x - \log M(\theta)] = \theta x - \log[\sum_{k=1}^{N} e^{\theta x_k} p_k],$$

We can get the probability distribution p_k of ship deformation according to statistical theories, we can calculate the speed of convergence

$$I(x) = \theta x - \log[\sum_{k=1}^{N} e^{\theta x_k} p_k] \tag{4.1}$$

Differentiate $I(x.\theta)$ with respect to θ, we have

$$\frac{\partial}{\partial\theta} I(x.\theta) = x - \frac{\sum_{k=1}^{N} x_k e^{\theta x_k} p_k}{\sum_{k=1}^{N} e^{\theta x_k} p_k} = 0$$

V. SIMULATION AND DISCUSS

Choose measured wave heights (Fig.1). From the heights in Fig.1, we can get the wave ' s spectral density with MATLAB ' s Fourier transform module (Fig.2). Set and , then we can get the probability distribution of ship deformation (Fig.3). At last, we can get the convergence function (Fig.4).

VI. CONCLUSIONS

It is of significance to study the accumulative deformation as the long ship deformation could lead to fatigue failure. This paper proves the convergence theory of large deviation according to ship deformation ' s time series. Moreover, calculate the convergence speed with respect to generating function and the ship deformation. The results provide a valuable reference to practice.

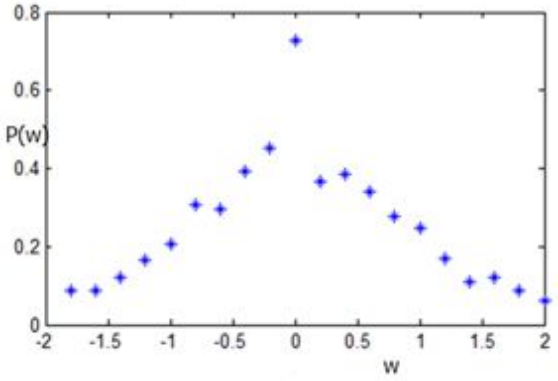


Fig.3 Probability distribution of ship deformation

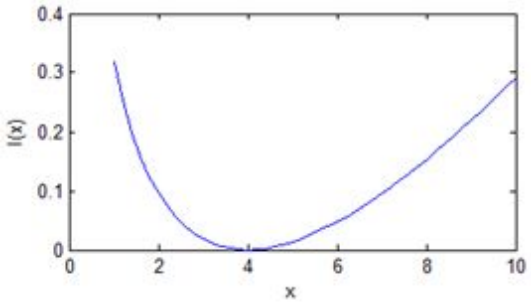


Fig.4 Convergence function

ACKNOWLEDGMENT

The paper is financially supported by China Ministry of Transport Applied Basic Research Project(No。 2013-329-811-360).

REFERENCES

[1] Li, T. Q., Shi, F. L., Liu, Z. Y., Chang, X., Price, W. G. and Temarel, P (2012). "Numerical simulation of multibody water impact using a two-phase solver", 22th International Offshore and Polar Engineering Conference (ISOPE), Vol. 4, pp 776-783. (EI)

[2] Zhixiong Li, Xinping Yan, Zhe Tain, Chengqing Yuan, Zhongxiao Peng. Blind vibration component separation and nonlinear feature extraction applied to the nonstationary vibration signals for the gearbox multi-fault diagnosis, Measurement, Elsevier。 2012.06.013.

[3] Zhixiong Li, Xinping Yan, Zhiwei Guo, Yuelei Zhang, Chengqing Yuan, Z. Peng. Condition Monitoring and Fault Diagnosis for Marine Diesel Engines using Information Fusion Techniques, Elektronika ir elektrotechnika, Kaunas: Technologija, 2012, 7(123): 109–112..

[4] Xin Chen, XinCong Zhou, Shesheng Zhang, Dan Li, A numerical fluid-solid coupling model for the dynamics of ships in atrocious sea conditions [J], Journal of Algorithms & Computational Technology,2015,Vol. 9 No. 2, 163-175.

[5] Li,5.Y.,Zhang,C.S.,WU,R.,Ruin estimates of diffusion models under constant interest rate.Chinese Jouxnal of Applied Probbaility and Statisties,2003,19(1),79 — 84.

[6] [6]Liao Yulin, Large deviations theory and application, Journal of Changsha railway institute,1987(01), 15-25..

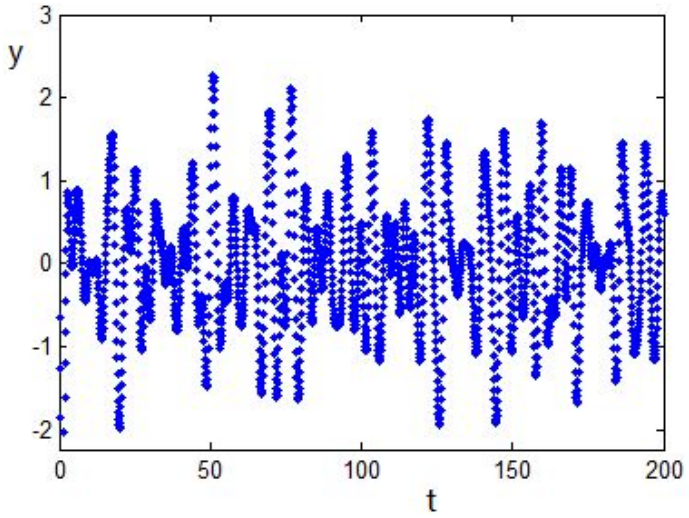


Figure 1. Measured wave height.

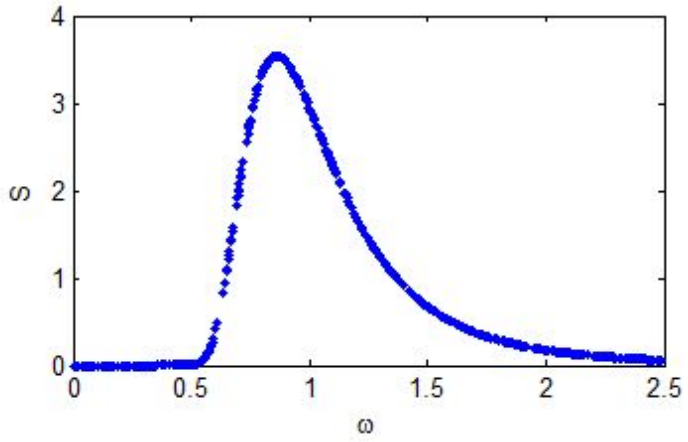


Figure 2. Wave’s spectral density

Distributed Adaptive Control of Diffusion System based on Multi-agents

Tiane Chen
School of IoT Engineering
Jiangnan university
WuXi,China
57458048@qq.com

Baotong Cui
School of IOT Engineering
Jiangnan university
WuXi,China
btcui@jiangnan.edu.cn

Zaihe Cheng
School of IOT technology
Wuxi Institute of technology.
WuXi,China
chengzh@wxit.edu.cn

***Abstract*—This paper addresses the problem of adaptive stabilization control and consensus control of diffusion system based on mobile multi-agents. Multi-agents triggered by events are integrated into diffusion system using self-measurement and neighbors information. Lyapunov redesign method is used to analyze adaptive feedback gain and consensus gain, and induce every agent trajectory. Then automatic guidance strategy is further enhanced to eliminate local falling problem and potential oscillatory behavior. Numerical simulation results show that overall performance can be achieved with the guidance and control strategy in the case of multiple static disturbances or mobile disturbance.**

Keywords- Multi-agents; Adaptive; Diffusion system; Guidance

I INTRODUCTION

In the field of environment protection, the pollution taking place should be tracked to ensure pollution concentration meet environmental standards. One effect control method is using multi-agents, where every agent consists sensor, actuator, controller, transceiver and mobile vehicle with sensor apperceiving environment variable, controller processing information, actuator driving the motion of mobile vehicle, and transceiver communicating with outside. Multi-agents' coupling degree is determined by communication and perception ability of every agent. Normally, agent works in the mode of cruising; Once pollution events happening, multi-agents are triggered into pollution diffusion system in the mode of control and navigated to coordinative control diffusion system.

Multi-agents triggered numerous studies in science and engineering. Many studies have focused on inter-coupling, time delay, regulation [1], consensus [2], tracking [3], and event driven [4]. In the majority of cases, multi-agents are not integrated into the controlled system. [5] leads remote control of UAV to distributed parameter system. [6] imported non-linear observe model. [7] introduced adaptive model reduction. [8] considered switch dynamic observer. Notable exception is the pioneering work [9], and automatic guidance trajectory is realized on the basis of system stability analysis. Stabilization and consensus control are considered in [10]. [11] studied consensus control depending on multi-agents network topology without achieving truly distributed control.

On the basis of above study[9-11], this paper proposes a fully distributed adaptive control algorithm, which doesn't need the network topology information, only using self-measurement and neighbors' information. Lyapunov redesign parameter function method is adopted twice to analyze stability. Gain of stabilization control and consensus control has been obtained. Then agent trajectory is induced using gain information known, finally agent moving velocity is optimized for eliminating local falling problem and oscillatory behavior.

II ABSTRACT FRAMEWORK AND PROBLEM STATEMENT

A. Diffusion system definition

Dynamic equation of pollution diffusion system is defined as:

$$\begin{cases} \frac{\partial x(t,\xi)}{\partial t} = a\frac{\partial^2 x(t,\xi)}{\partial \xi^2} + \mathrm{d}(\mathrm{t},\xi) \\ x(0,\xi) = x_0(\xi) \end{cases} \quad (1)$$

Where a is system diffusion coefficient, $x(t,\xi)$ represents system state at time t and position ξ, $x_0(\xi)$ is system initial state, $\mathrm{d}(\mathrm{t},\xi)$ represents disturbance.

B. Multi-agents control

There are N agents, each agent binds sensor and actuator, each agent is assumed to be massless and inertialess, and each agent can intercommunicate with self-neighbors.

1) Spatial distribution:

Sensors are point-wise with spatial distribution being δ function defined as:

$$\mathrm{h} = \delta(\xi_i^{\,s}(t)) \quad (2)$$

Where $\xi_i^{\,s}(t)$ is mass centroid of sensor i. Actuators are point-wise with spatial distribution being δ function:

$$\mathrm{g} = \delta(\xi_i^{\,a}(t)) \quad (3)$$

Where $\xi_i^{\,a}(t)$ is mass centroid of actuator i [13]. It's assumed that actuators and sensors in the network are collocated, meaning that:

$$\xi_i^{\,s}(t) = \xi_i^{\,a}(t) = \xi_i(t) \quad (4)$$

2) Working mode:

Agent is working in the event-driving mode. When every neighbor agent measurement y_j is less than threshold value μ, and current agent measurement value y_i is less than μ, then agent works in the mode of cruising according to predetermined trajectory with only detecting, given by: $y_j < \mu$ and $y_i < \mu$. Otherwise, multi-agents are integrated into diffusion system, and agent is working in controlling mode with automatic navigation and pollution control.

C. Controller design

1) Control objective

Control objective consists of feedback control and consensus control. The control objective is to choose the control signals, so that all states and all pairwise differences asymptotically converge to zero.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.109

Feedback control objective：

$$lim_{t\to\infty} x_i(t)=0 \quad (5)$$

Consensus control objective：

$$lim_{t\to\infty}[x_i(t)-x_j(t)]=0 \quad (6)$$

Because system state can't be easily obtained, aforementioned formula can be converted to output consensus control objective defined as:

$$lim_{t\to\infty}[y_i(t)-y_j(t)]=0 \quad (7)$$

2) Control system framework

Static output feedback control is chosen for its important features, such as simple structure, least computation complexity and fast response speed. This framework can reduce the effect of time delay, and feedback output control can be represented as: $u_{i0}=-\delta_i y_i$, where δ_i denots feedback gain. Every agent utilizes neighbor's information, so consensus control can be represented as:

$$\mathrm{u_{i1}}=-\frac{1}{Ni}\sum_{j=0}^{Ni}\gamma_{ij}(y_i-y_j) \quad (8)$$

γ_{ij} is consensus gain. When current agent i measurement y_i or neighbor's measurement y_j is over threshold μ, system input control is valid, and defined as：

$$u_i=\begin{cases}-\delta_i y_i-\frac{1}{Ni}\sum_{j=0}^{Ni}\gamma_{ij}(y_i-y_j), & y_i>\mu \text{ or } y_j>\mu\\ 0, & y_i\le\mu\end{cases} \quad (9)$$

Controller design and stability analysis are always involved, so the control gain are realized using stability analysis method in the third section.

3) Control system model

Rewrite multi-agents control system model:

$$\begin{cases}\frac{\partial x_i(\xi,\ \xi_i\ (t))}{\partial t}=a\frac{\partial^2 x_i(\xi,\ \xi_i\ (t))}{\partial\xi^2}+g(\xi,\xi_i^a(t))u_i(t)\\ u_i=-\delta_i y_i-\frac{1}{Ni}\sum_{j=0}^{Ni}\gamma_{ij}\left(y_i-y_j\right)\\ y_i(t)=\int_0^l h(\xi,\xi_i^s(t))x(t,\xi)\,d\xi\\ g\left(\xi,\xi_i^s(t)\right)=h(\xi,\xi_i^a(t))=\begin{cases}1 & \xi=\xi_i(t)\\ 0 & \xi\ne\xi_i(t)\end{cases}\end{cases} \quad (10)$$

Where the Hilbert space$\{\mathcal{H},<\cdot,\cdot>_{\mathcal{H}},|\cdot|_{\mathcal{H}}\}$ is the state space. The space$\{\mathcal{V},|\cdot|_{\mathcal{V}}\}$is a reflexive Banach space densely and continuously embedded in $\mathcal{H}$,$\mathcal{V}^*$ denotes the continuous dual of $\mathcal{V}$. Then $\mathcal{V}$->$\mathcal{H}$->$\mathcal{V}^*$[7], The state operator $\mathcal{A}$: $\mathcal{V}$ -> $\mathcal{V}^*$ is a boundness, coercivity and symmetry linear operator, with $\mathcal{A}\phi=\frac{\mathrm{d}}{\mathrm{d}\xi}(\mathrm{a}\frac{\mathrm{d}\phi}{\mathrm{d}\xi})$, a >0, Dom($\mathcal{A}$) is square integrable. According to operator semi-group theory, $\mathcal{A}$ generates an exponentially stable C_0 semi-group, with the property:

$$\mathcal{A}+\mathcal{A}^*\le KI \quad (11)$$

Diffusion control system evolution model can be rewrite as following:

$$\dot{x_i}=\mathcal{A}x(t)-G_i(\xi_i^a)\delta_i y_i(t)-G_i(\xi_i^a)\gamma_i(y_i-\bar{y_j}) \quad (12)$$

And input operator is given by

$$<G_i(\xi_i^a)u_i(t),\phi>=\int_0^l g_i(\xi_i^a)\phi(\xi)u_i(t)\,d\xi \quad (13)$$

III Analyzing adaptive Adjustment rate

Lemma1: Multi-agents control system model satisfies (10), Agent's sensor and actuator are collocated, and their distributions are δ function of mass centroid. Multi-agents control system is stabilized with adaptive adjustment rate of feedback gain and consensus gain given by:

$$\dot{\delta_i}=y_i(t)^2-\sigma_{01}\delta_i,\ \dot{\gamma_{ij}}=y_i(y_i-y_j)-\sigma_{02}\gamma_{ij} \quad (14)$$

Where $\sigma_{01}>0$, $\sigma_{02}>0$, i=1，2，……N

Proof: Select Lyapunov function $V_1(t)$:

$$V_1(t)=<x_i(\xi,t),x_i(\xi,t)>+<\delta_i,\delta_i>+<\gamma_{ij},\gamma_{ij}> \quad (15)$$

The derivative of the Lyapunov function is:

$$\begin{aligned}\dot{V}_1(t)=&<\mathrm{x_i}(\dot{\xi},t),\mathrm{x_i}(\xi,t)>+<\mathrm{x_i}(\xi,t),\dot{\mathrm{x}}_\mathrm{i}(\xi,t)>\\ &+2<\dot{\delta}_\mathrm{i},\delta_\mathrm{i}>+2<\dot{\gamma}_{\mathrm{ij}},\gamma_{\mathrm{ij}}>\\ =&<\mathcal{A}\mathrm{x_i}(\xi,t),\mathrm{x_i}(\xi,t)>+<\mathrm{x_i}(\xi,t),\mathcal{A}\mathrm{x_i}(\xi,t)>\\ &+<-\mathrm{G_i}(\xi_\mathrm{i}^\mathrm{a})\delta_\mathrm{i}\mathrm{y_i}(t),\mathrm{x_i}(\xi,t)>\\ &+<\mathrm{x_i}(\xi,t),-\mathrm{G_i}(\xi_\mathrm{i}^\mathrm{a})\delta_\mathrm{i}\mathrm{y_i}(t)>+2<\dot{\delta}_\mathrm{i},\delta_\mathrm{i}>\\ &+\tfrac{1}{\mathrm{Ni}}\textstyle\sum_{\mathrm{j}=0}^{\mathrm{Ni}}<-\mathrm{G_i}(\xi_\mathrm{i}^\mathrm{a})\gamma_{\mathrm{ij}}(\mathrm{y_i}-\mathrm{y_j}),\mathrm{x_i}(\xi,t)>\\ &+\tfrac{1}{\mathrm{Ni}}\textstyle\sum_{\mathrm{j}=0}^{\mathrm{Ni}}<\mathrm{x_i}(\xi,t),-\mathrm{G_i}(\xi_\mathrm{i}^\mathrm{a})\gamma_{\mathrm{ij}}(\mathrm{y_i}-\mathrm{y_j}),\mathrm{x_i}>\\ &+2<\dot{\gamma}_{\mathrm{ij}},\gamma_{\mathrm{ij}}>\end{aligned}$$

Decompose the derivative function $\dot{V_1}(t)$ into two sub functions $\dot{V_{11}}$ and $\dot{V_{12}}$, then:

$$\begin{aligned}\dot{V_{11}}=&<Ax_i(\xi,t),x_i(\xi,t)>+<x_i(\xi,t),\mathcal{A}x_i(\xi,t)>\\ =&<(\mathcal{A}+\mathcal{A}^*)x_i(\xi,t),x_i(\xi,t)>\\ \le&-k||x_i(\xi,t)||^2\end{aligned}$$

$$\begin{aligned}\dot{V_{12}}=&<-G_i(\xi_i^a)\delta_i y_i(t),x_i(\xi,t)>\\ &+<x_i(\xi,t),-G_i(\xi_i^a)\delta_i y_i(t)>+2<\dot{\delta_\imath},\delta_i>\\ =&\,2<\delta_i,-G_i(\xi_i^a)x_i(\xi,t)y_i(t)+\dot{\delta_\imath}>\end{aligned}$$

Select $-G_i(\xi_i^a)x_i(\xi,t)y_i(t)+\dot{\delta_\imath}=-\sigma_{01}\delta_i$

where $\sigma_{01}\ge 0$, and $\dot{\delta_i}=y_i(t)^2-\sigma_{01}\delta_i$, then:

$$\dot{V_{12}}=2<\delta_\mathrm{i},-\sigma_{01}\delta_\mathrm{i}>=-\sigma_{01}||\delta_\mathrm{i}||^2$$

$$\begin{aligned}\dot{V_{13}}=&\tfrac{1}{Ni}\textstyle\sum_{j=0}^{Ni}<-G_i(\xi_i^a)\gamma_{ij}(y_i-y_j),x_i(\xi,t)>\\ &+\tfrac{1}{Ni}\textstyle\sum_{j=0}^{Ni}<x_i(\xi,t),-G_i(\xi_i^a)\gamma_{ij}(y_i-y_j),x_i>\\ &+2<\dot{\gamma_{ij}},\gamma_{ij}>\\ =&2*(\tfrac{1}{Ni}\textstyle\sum_{j=0}^{Ni}<\gamma_{ij},-G_i(\xi_i^a)x_i(\xi,t)(y_i-y_j)+\dot{\gamma_{ij}}>)\end{aligned}$$

Select$-\mathrm{G_i}(\xi_\mathrm{i}^\mathrm{a})\mathrm{x_i}(\xi,\mathrm{t})(\mathrm{y_i}-\mathrm{y_j})+\dot{\gamma}_{\mathrm{ij}}=-\sigma_{02}\gamma_{\mathrm{ij}}$,$\sigma_{02}\ge 0$

$\dot{\gamma}_{\mathrm{ij}}=\mathrm{G_i}(\xi_\mathrm{i}^\mathrm{a})\mathrm{x_i}(\xi,\mathrm{t})(\mathrm{y_i}-\mathrm{y_j})-\sigma_{02}\gamma_{\mathrm{ij}}=\mathrm{y_i}(\mathrm{y_i}-\mathrm{y_j})-\sigma_{02}\gamma_{\mathrm{ij}}$,

Then $\dot{V_{13}}=2*\left(\frac{1}{Ni}\sum_{j=0}^{Ni}<\gamma_{ij},-\sigma_{02}\gamma_{ij}>\right)=-2*\sigma_{02}||\gamma_{ij}||^2$

Provided feedback control gain and consensus control gain will not be zero at the same time, then $\dot{V_1}(t)$ is less than zero, and system is stabilized, and lemm1 is proved. From the lemma 1, feedback control gain is related to current agent measurement, consensus control gain is relevant to current measurement and neighbor's information. Each agent implement fully distributed control, without considering the knowledge of multi-agents topology, using only local information. Further select consensus gain:

$\dot{\gamma_\imath}=\frac{1}{Ni}\sum_{j=0}^{Ni}\dot{\gamma_{ij}}=-\mathrm{y_i}(y_i-\bar{y_j})-\sigma_{02}\frac{1}{Ni}\sum_{j=0}^{Ni}\gamma_{\mathrm{ij}}$,

That's $\dot{\gamma_\imath}=y_i(y_i-\bar{y_j})-\sigma_{02}\gamma_i$, Then:

$$u_i=-\delta_i y_i-\gamma_i(y_i-\bar{y_j}) \quad (16)$$

Consensus gain is greatly simplified and Every agent utilizes self-information and neighbors average fuse information. At the same time an alternate consensus control objective is reformed as the deviation from the mean.

$$lim_{t\to\infty}[y_i(t)-\tfrac{1}{Ni}\textstyle\sum_{j=0}^{Ni}y_j(t)]=lim\,[_{t\to\infty}y_i(t)-\bar{y_j}(t)]=0 \quad (17)$$

IV BACKSTEPPING AGENT TRAJECTORY DESIGN

A. Agent trajectory

Lemma2: Diffusion control system model satisfies equation (10). Agent's sensor and actuator are collocated, and their distributions are δ function of mass centroid. Sufficient conditions for Multi-agents control system stability is given by:

$$\dot{\xi}_i^a = \sigma y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right) - \frac{a}{2} \tag{18}$$

where $\sigma > 0$, $\dot{\xi}_i^a$ is agent's velocity, $y_{\xi_i^a}^{\cdot}$ denotes spatial derivative of measurement at position ξ_i^a. a is diffusion coefficient. δ_i and γ_i represent feedback gain and consensus gain respectively.

Proof: Select Lyapunov function $V_2(t)$:

$$V_2(t) = -< x_i(\xi,t), \dot{x}_i(\xi,t) > \tag{19}$$

The derivative of the Lyapunov function is:

$$\begin{aligned}\dot{V}_2(t) = &-< \dot{x}_i(\xi,t), \dot{x}_i(\xi,t) > -< x_i(\xi,t), A\dot{x}_i(\xi,t)\\ &-\dot{G}_i(\xi_i^a)\delta_i y_i(t) - G_i(\xi_i^a)\dot{\delta}_i y_i(t) - G_i(\xi_i^a)\delta_i \dot{y}_i(t)\\ &-\dot{G}_i(\xi_i^a)\gamma_i\left(y_i - \overline{y}_j\right) - G_i(\xi_i^a)\dot{\gamma}_i\left(y_i - \overline{y}_j\right)\\ &-G_i(\xi_i^a)\gamma_i\left(\dot{y}_i - \dot{\overline{y}}_j\right) >\end{aligned}$$

Decompose the derivative function into two sub functions :

$$\begin{aligned}\dot{V_{21}}(t) = &-< x_i(\xi,t), -\dot{G}_i(\xi_i^a)\delta_i y_i(t) - G_i(\xi_i^a)\delta_i \dot{y}_i(t)\\ &-\dot{G}_i(\xi_i^a)\gamma_i\left(y_i - \overline{y}_j\right) - G_i(\xi_i^a)\gamma_i\left(\dot{y}_i - \dot{\overline{y}}_j\right) >\\ = &\ -2\dot{\xi}_i^a y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right)\end{aligned}$$

$$\begin{aligned}\dot{V_{22}}(t) = &-< \dot{x}_i(\xi,t), \dot{x}_i(\xi,t) > -< x_i(\xi,t), A\dot{x}_i(\xi,t) -\\ &G_i(\xi_i^a)\dot{\delta}_i y_i(t) - G_i(\xi_i^a)\dot{\gamma}_i\left(y_i - \overline{y}_j\right) >\\ = &-< \dot{x}_i(\xi,t), \dot{x}_i(\xi,t) > -< Ax_i(\xi,t), Ax_i(\xi,t)\\ &-G_i(\xi_i^a)\delta_i y_i(t) - G_i(\xi_i^a)\gamma_i\left(y_i - \overline{y}_j\right) >\\ &-< x_i(\xi,t), G_i(\xi_i^a)\dot{\delta}_i y_i(t) + G_i(\xi_i^a)\dot{\gamma}_i\left(y_i - \overline{y}_j\right) >\end{aligned}$$

Select $\dot{\gamma}_i = y_i\left(y_i - \overline{y}_j\right)$ and $\dot{\delta}_i = y_i(t)^2$, then:

$$\begin{aligned}\dot{V}_2(t) = &-2\dot{\xi}_i^a y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right) -< \dot{x}_i(\xi,t), \dot{x}_i(\xi,t)\\ &> -< x_i(\xi,t), Ax_i(\xi,t) > -< {y_i}^2, {y_i}^2 > -\\ &< {y_i}^2, \left(y_i - \overline{y}_j\right)^2 > \ -a\, y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right)\end{aligned}$$

Let: $-\left(2\dot{\xi}_i^a + a\right) y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right) = -\sigma' {y_{\xi_i^a}^{\cdot}}^2$

$(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right))^2$, and $\dot{\xi}_i^{'a} = \dot{\xi}_i^a + \frac{a}{2}$, $\sigma = \frac{\sigma'}{2} > 0$.

Select: $\dot{\xi}_i^{'a} = \sigma y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right)$, That is :

$$\dot{\xi}_i^a = \sigma y_{\xi_i^a}^{\cdot}\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right) - \frac{a}{2}$$

Provided current measurement and the deviation from the mean are not zero at the same time, then $\dot{V}_2(t)$ is less than zero, and system is stable. Lemm2 is proved. From the lemma 2, agent velocity's orient is system state derivative direction at current position. And σ is constant greater than zero. Oscillatory behavior may occur when derivative is sudden change. Local falling problem can also happen, leading to insufficient global impact.

B. Optimization of agent trajectory

Aiming at above problem from agent moving velocity, agent trajectory joins two stage cycle model of bird foraging [14]. The bird forages food nearby in the inner cycle and flies to the location of a large number of food. Assume that agent broadcasts self-measurement and self-position[15]. Agent stands still in the case of $(y_i - \overline{y}_j) \geq 0$, and agent plays mobile and control role in the case of $(y_i - \overline{y}_j) < 0$.

Current agent i searches the neighbor which measurement is greater than average measurement within the radius of average neighbors' position. If neighbor j0 satisfying is found, then agent executes inner loop with moving orient determined by $\left(y_i - y_{j0}\right)\left(\xi_i - \xi_{j0}\right)$ and actual value of $\sigma y_{\xi_i^a}^{\cdot}$ selected as:

$$k1 = \sigma'\left(y_i - y_{j0}\right)\left(\xi_i - \xi_{j0}\right) \tag{20}$$

Otherwise agent executes outer loop with moving orient determined by $(y_i - \overline{y}_j)(\xi_i - \overline{\xi}_j)$ .And actual value of $\sigma y_{\xi_i^a}^{\cdot}$ replaced by:

$$k2 = \sigma''\left(y_i - \overline{y}_j\right)\left(\xi_i - \overline{\xi}_j\right) \tag{21}$$

Agent trajectory of mobile control is defined as :

$$\dot{\xi}_i^a = \begin{cases}\sigma' k1\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right) - \frac{a}{2} & Constr\\ \sigma'' k2\left(\delta_i y_i(t) + \gamma_i\left(y_i - \overline{y}_j\right)\right) - \frac{a}{2} & Otherwise\end{cases} \tag{22}$$

Where $(y_i - \overline{y_j}) < 0$, and Constr=$y_{j0} > \overline{y}_j$ & $|\xi_i - \xi_{j0}| < |\xi_i - \overline{\xi}_j|$, and $\overline{\xi}_j = \sum_1^{Ni} \xi_j$. When $(y_i - \overline{y_j}) \geq 0$,agent keeps still and $\dot{\xi}_i^a = 0$.

Because agent navigation speed is limited by maximum speed, agent i's speed satisfies the following formula:

$$v_i(t) = \begin{cases}\dot{\xi}_i^a & , abs(\dot{\xi}_i^a) < v_{max}\\ sgn(\dot{\xi}_i^a) * v_{max} & , abs(\dot{\xi}_i^a) \geq v_{max}\end{cases} \tag{23}$$

Where sgn is sign function, v_{max} represents maximum speed.

Remark: The closer the distance between the agent and the target, the smaller the speed. The smaller the measurement difference, the smaller the speed. This improved scheme can eliminate oscillatory behavior. The dual cycle not only considers the influence of the agent to the global, but also considers the process of the agent to the local.

V EXAMPLE

The coefficient of diffusion operator in the diffusion control system (1) was $\boldsymbol{a}$=0.00001, $\xi \in [0,\ \ell]$, ℓ =1, i=1,2,⋯⋯5。

Agents' initial position（ξ_1、ξ_2、ξ_3、ξ_4、ξ_5）is respectively（0.1 ℓ, 0.3 ℓ, 0.5 ℓ, 0.7 ℓ, 0.9 ℓ）. Initial static pollution sources are: x_1、x_2、x_3、x_4、x_5, and their pollution values are:

$$\begin{cases}x_1(0, \xi_1 - 0.04\ell) = 1.4cos\ (1.3\pi\xi_1)\\ x_2(0, \xi_2 - 0.04\ell) = 8.6\ cos(2.1\pi\xi_2)\\ x_3(0, \xi_3 + 0.04\ell) = 7.6cos\ (1.6\pi\xi_3)\\ x_4(0, \xi_4) = 7.5cos\ (4.7\pi\xi_4)\\ x_5(0, \xi_5) = 6.2(cos(5.1\pi\xi_5) + 1.3)\end{cases}$$

In the first simulation, There are three cases. The first case is open loop system without agent. Agent is integrated into system and plays feedback control role in the second case. Agent is integrated into system considering feedback control and consensus control in the third case. Simulation results are shown as Figure1。

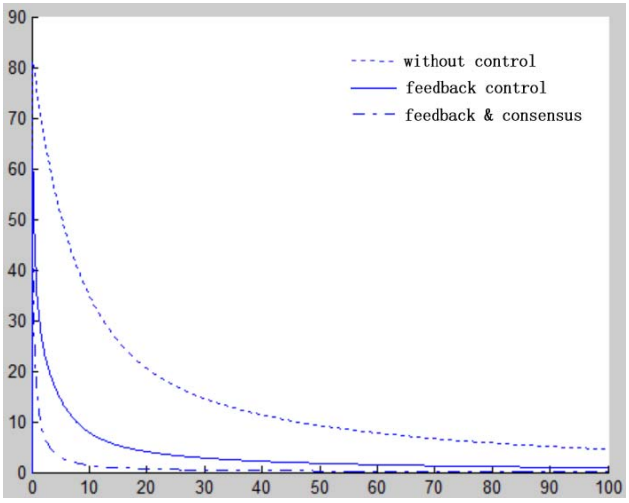


Figure 1 three case comparison of total state s uare erro

From the simulation, we can see that system energy with feedback and consensus control can converge more quickly and system square error also converge faster.

Taking the same mobile disturbance source $d(t,\xi)$ as [9] into system with position $\xi_d(t)$:

$$d(t,\xi) = 10^{-6}(3\,cos\left(\frac{9\pi t}{t_f}\right) + 5), \xi_d(t) = 0.2 * cos\left(\frac{9\pi t}{20}\right) + 0.7$$

Adaptive control simulations are executed in the case of mobile agent and static agent. As shown in the figure2, solid line represents mobile agent case and dot line denotes static agent case.

From this simulation, we can see that mobile agents control system energy with feedback and consensus control has faster convergence than static agents do.

VI CONCLUSION

Distributed adaptive control of diffusion system based on multi-agents improve system control performance by feedback control and consensus control. System output and consensus converge more quickly through adaptively changing feedback gain and consensus gain and event-driven agents' trajectory. This adaptive strategy can be extended to extensive application areas, and further study is time delay and formation control of multi-agents.

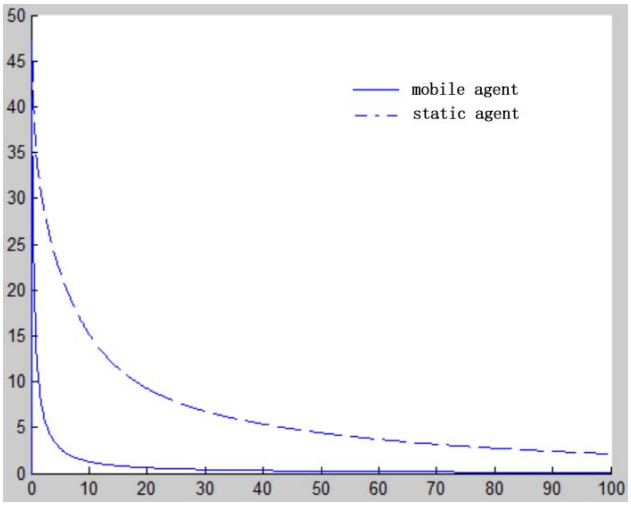


Figure2 the comparison of total sum of squared difference in the case of mobile and static system

REFERENCES

[1] C Huang, X Ye, Z Sun , Output regulation problem of multi-agents in networked systems , IET control theory & applications, 2012

[2] A Garulli, A Giannitrapani , Analysis of consensus protocols with bounded measurement errors, Systems & Control Letters, 2011

[3] R Vatankhah, S Etemadi, A Alasty, G Vossoughi , Adaptive critic-based neuro-fuzzy controller in multi-agents: Distributed behavioral control and path tracking, Neurocomputing, 2012

[4] Y Fan, G Feng, Y Wang, C Song , Distributed event-triggered control of multi-agent systems with combinational measurements , Automatica 49 (2013) 671–675

[5] H Chao, YQ Chen , Remote sensing and actuation using unmanned vehicles, books.google.com.sci-hub.org, 2012

[6] S Kar, JMF Moura, K Ramanan, Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 58, NO. 6, JUNE 2012

[7] S Pitchaiah, A Armaou, Output feedback control of dissipative PDE systems with partial sensor information based on adaptive model reduction, AIChE Journal, 2013

[8] DB Pourkargar, A Armaou, Design of APOD-based switching dynamic observers and output feedback control for a class of nonlinear distributed parameter systems, Chemical Engineering Science, 2015

[9] Michael A.Demetriou ,Guidance of Mobile Actuator-Plus-Sensor Networks for Improved Control and Estimation of Distributed Parameter System, IEEE Transactions on automatic control,vol.55,No.7,July 2010

[10] Michael A.Demetriou ,Adaptation and optimization of synchronization gains in the regulation control of networked distributed parameter systems,IEEE Transactions on Automatic Control ,2014

[11] Michael A.Demetriou ,Adaptive output feedback synchronization of networked distributed parameter systems,2014 American Control Conference(ACC) June 4-6,2014

[12] Nojeong Heo and P.K.Varshney, Energy-efficient deployment of intelligent mobile sensor networks , IEEE Transactions on Man and Cybernetics systems, vol. Part A,pp.78-92,January 2005

[13] C Tricaud, YQ Chen ,Optimal Mobile Sensing and Actuation policies in Cyber Physical Systems, Springer,2012

[14] Huibin Chang, Danping Yang, A Schwarz domain decomposition method with gradient projection for optimal control governed by elliptic partial differential equations, Journal of Computational and Applied Mathematics 235 (2011) 5078–5094

[15] B Stark, S Rider, YQ Chen, Optimal control of a diffusion process using networked unmanned aerial systems with smart health, World Congress, 2014

Performance analysis and simulation of vehicle electronic stability control system

Yang Ying[1], Liu Weiguo[2], Isah Sagir Tukur[1]

1.School of Mechanical Engineering and Automation
Northeastern University
Shenyang, China
yangyang@mail.neu.edu.cn

2. Zhejiang Key Laboratory of Automobile Safety Technology, Hangzhou, China
Hangzhou, China
lwg@rd.geely.com

***Abstract*—The benefits of Vehicle Electronic Stability Control (ESC) are well understood with regard to assisting drivers to maintain vehicle control during extreme handling maneuvers or when extreme road conditions are encountered. The goal of this paper is to study and develop an Electronic Stability Control (ESC) system model by utilizing the functionality of Matlab-Simulink and used Carsim to validate the developed ESC model. A certain vehicle dynamics model created in CarSim was validated by comparing simulation results with quasi-static and dynamic test data. The ESC system model was built in Simulink and was tuned through cosimulation with the CarSim vehicle model. The ESC system was designed to operate in Yaw Stability Control (YSC) and Roll Stability Control (RSC) modes. The performance of the system was evaluated using the Sine with Dwell and Fishhook maneuvers. This study proved the possibility to develop functional ESC system given a Carsim vehicle model and simulate the performance comparable to actual ESC system onboard the vehicle.**

Keywords- ESC; Carsim; Vehicle dynamic; Simulation

I. INTRODUCTION

A healthy transportation system and a nation's economy compliment and support each other. China has the 2nd highest national GDP, has a staggering number of vehicles on its roads. The number of crashes and injuries reported annually is also very high. Automotive safety has been a critical area for all automotive companies. Automotive passive safety is very beneficial in frontal impact crashes but less effective in side impact and rollovers [1]. Active safety systems have received a lot of attention in the last few years. These systems can help in preventing accidents as they attempt to keep the vehicle within its stability envelope. For example, the vehicle stability control is the Electronic Stability Control (ESC) system [2].

The objective of this study is to build an ESC model that will give comparable performance to actual ESC systems on board the vehicle for certain maneuvers. The utility of such a model is that it can further be used to test the effect of new components or systems (for example an active suspension system) on an ESC-equipped vehicle. In such studies, it is not prudent to spend time and resources in building a comprehensive ESC model. A model, as described in this study, has considerably less development time and still gives comparable performance to the actual ESC system, in the specific maneuvers. This allows us to allocate resources to the main task of evaluating the effect of the new component or system on the performance of the ESC-equipped vehicle.

II. VEHICLE MODELING

Wherever CarSim is commercially available vehicle dynamics simulation software. It has a large database made up of more than 150 libraries of datasets linked together [3-5]. These datasets contain vehicle model parameters and simulation settings. Datasets of a generic SUV in CarSim were suitably modified to build a certain SUV models. The vehicle modeled is a 4-door sports utility vehicle with a 5.4L V8 engine, automatic transmission, 4WD and P265/70R17 113S M+S tires. Tire pressures are 35 psi for both front and rear tires.

The two track model is an extension of the single track model. It has two axles and four wheels with a width w between the wheels of the same axle. From figure 1, the x and y directions can again be described as the longitudinal and lateral directions respectively. All other symbols are as defined for the bicycle model.

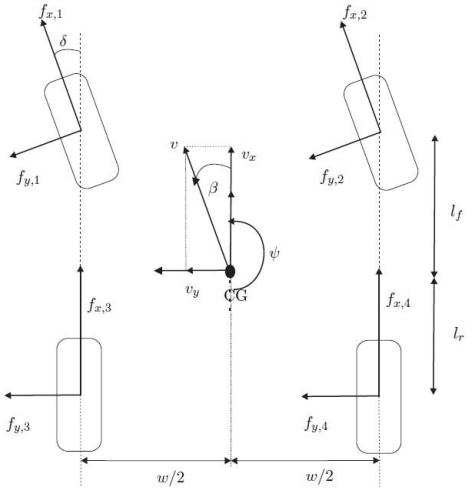


Figure 1. Two Track Vehicle Model

For each of the four wheels in the two track model, the contribution of both longitudinal and lateral forces $f_{x,i}$ and $f_{y,i}$, can be translated into the vehicle axis system at the vehicle centre of gravity (CG), $F_{x,i}$ and $F_{x,i}$.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.110

$$F_{x,i} = f_{x,i}\cos\delta_i - f_{y,i}\sin\delta_i$$
$$F_{y,i} = f_{x,i}\sin\delta_i + f_{y,i}\cos\delta_i \quad (1)$$

To keep these expressions tidy, a Rotation Matrix, $D(\delta_i)$ is introduced:

$$D(\delta_i) = \begin{bmatrix} \cos\delta_i & -\sin\delta_i \\ \sin\delta_i & \cos\delta_i \end{bmatrix} \quad (2)$$

And so:

$$\begin{bmatrix} F_{x,i} \\ F_{y,i} \end{bmatrix} = D(\delta_i)\begin{bmatrix} f_{x,i} \\ f_{y,i} \end{bmatrix} \quad (3)$$

Applying Newton's second law of motion to the vehicle body, the following equations describe the forces at the vehicle centre of gravity in its own co-ordinate system.

$$m(\dot{v}_x - v_y\dot{\varphi}) = \sum_{i=1}^{4} F_{x,i}$$
$$m(\dot{v}_y + v_x\dot{\varphi}) = \sum_{i=1}^{4} F_{y,i} \quad (4)$$

Finally, the sum of the moments around the vehicle is equal to the moment around the centre of gravity,

$$J_z\ddot{\varphi} = \sum_{i=1}^{4} M_{z,i} \quad (5)$$

III. Electronic Stability Control Model

The ESC system model was built in Simulink and was tuned through cosimulation with CarSim. CarSim and Simulink interact through the CarSim S-function block. Parameters such as vehicle speed, steering wheel angle, lateral acceleration, yaw rate and longitudinal slip ratios are obtained from the CarSim vehicle model and are used as inputs to the ESC model. The steering input given to the vehicle model is essentially passed through the CarSim vehicle model and sent to the ESC system. The brake cylinder pressures calculated by the ESC model are sent to the CarSim model. The thresholds for the ESC system, namely values of lateral acceleration and vehicle slip rate are specified in the Matlab-Simulink environment. The memory element before the CarSim S-function introduces a delay of unit time step in the simulation. This leads to the parameters obtained from CarSim at a particular time step, to be used by the ESC model in the next time step. This in turn prevents the simulation from going into an infinite algebraic loop

A. *Roll Stability Control*

First, The Roll Stability Control module (shown in Fig. 2) contains the control logic for selecting the wheel to be braked and for calculating the magnitude of braking force, so that the vehicle does not rollover. Lateral acceleration is obtained from the CarSim vehicle model. The magnitude of braking force is calculated using a look-up table which uses lateral acceleration as input and gives brake pressure as the output. In this model, the look-up table uses an approximately linear relation between lateral acceleration and brake pressure. The saturation block limits the brake pressure to that allowed by the vehicle brake system. The time delay block is used to simulate the time delay in the brake system – from the time ESC is activated to the time brakes are actually applied. The brake selector block is a Matlab function which selects the wheel to be braked. The steering wheel angle obtained from CarSim is used to determine the direction of turn of the vehicle and accordingly, the outer front wheel is braked. The brake pressures calculated in this module are then passed to the CarSim vehicle model.

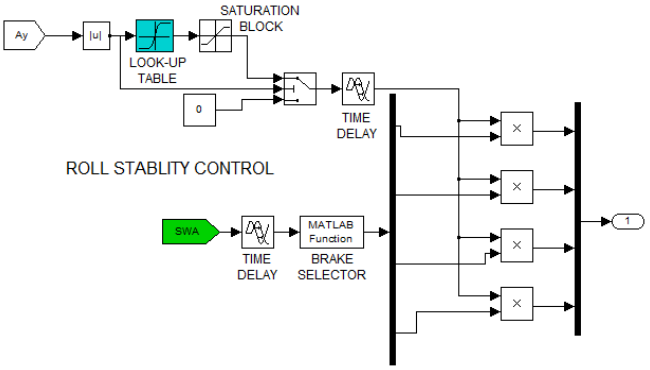


Figure 2. Roll Stability Control Module

B. *Yaw Stability Control*

This module (shown in Fig. 3) uses differential braking to ensure that the vehicle retains directional stability. The vehicle slip rate (difference between actual and ideal yaw rates) is obtained from the activation module.

The magnitude of this control variable determines the magnitude of the brake pressure and their relation is entered in the look-up table. As in the RSC module, the saturation block in Fig. 3 limits the brake pressure to the brake system capacity and the time delay block simulates the delay in the brake system. Vehicle slip rate, its threshold value and steering wheel angle are used as inputs to the brake selector block. The direction of turn and the sign of vehicle slip rate are used to select the wheel which is to be braked.

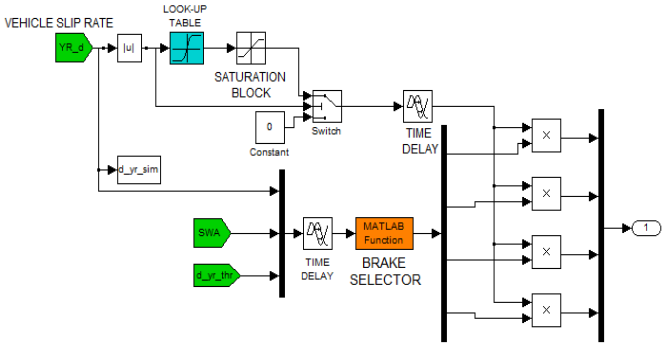


Figure 3. Yaw Stability Control Module

The threshold values of all control variables are specific to a vehicle and have to be determined through test data or through simulations (if a validated vehicle model is available). Thus for this model, there are three thresholds: (1) lateral acceleration threshold for RSC (2) vehicle slip rate threshold for YSC and (3) longitudinal slip ratio threshold for the ABS functionality.

IV. Simulation Results

The simulation results contain the co-simulation between the vehicle model in CarSim and the ESC model in Matlab-Simulink. The test conditions, model parameters and thresholds used for the tests are also described. The simulation results are then used to evaluate the performance of the ESC system model.

A. Yaw Stability Control Mode

The sine with dwell (SWD) test is used to evaluate the performance of yaw stability control mode of the ESC system. Fig. 4 shows the comparison plots of simulation results and test data – in the baseline mode (ESC turned OFF) and with ESC. They show that the actual and modeled vehicle - with ESC ON and OFF – had the same steering input (SWD of magnitude 160 degrees) and same initial speed (80 kmph). This particular steering input was chosen because the actual vehicle was found to spin out for this steering input during tests. In other words, this was a limit maneuver for the vehicle and the effect of ESC intervention is clearly seen in the lateral acceleration and yaw rate response plots. The test vehicle with ESC 'OFF' and the modeled baseline vehicle are both seen to spin out in this maneuver.

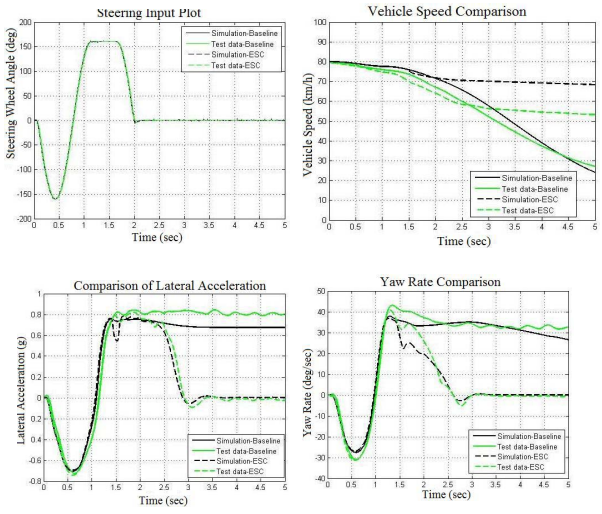


Figure 4. Effect of ESC Intervention

For the test and modeled vehicles with ESC 'ON', the lateral acceleration and yaw rate both return to zero which shows that ESC intervention prevents the vehicle from spinning out. Further, the performance of the ESC system model is comparable to the actual ESC system on board the vehicle, for this particular maneuver. The vehicle slip rate threshold used for the ESC model was 0.18.

The simulation results showed the improvement in lateral stability due to ESC by comparing the response of baseline and ESC equipped vehicles.

The set of Fig. 5 compare the performance of the ESC system model and the actual ESC-equipped vehicle in further detail.

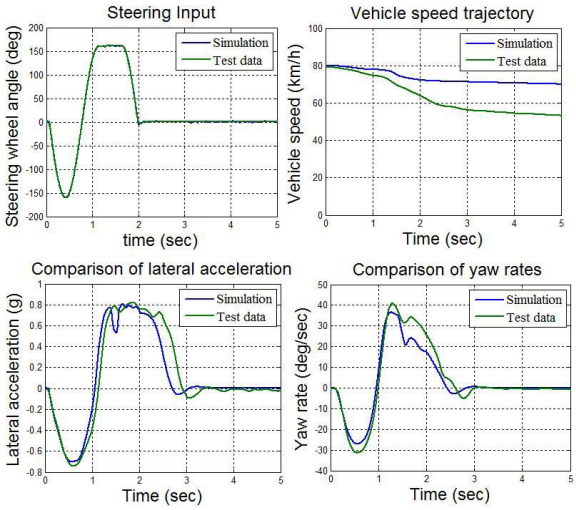


Figure 5. Sine with Dwell -160: Steering Input, Vehicle Speed, Lateral Acceleration and Yaw Rate

The result shows that the test and simulated vehicles have the same steering input and initial speed. Further the lateral acceleration and yaw rate response plots show that the performance of the ESC system model and actual on-board ESC system is also comparable for this maneuver.

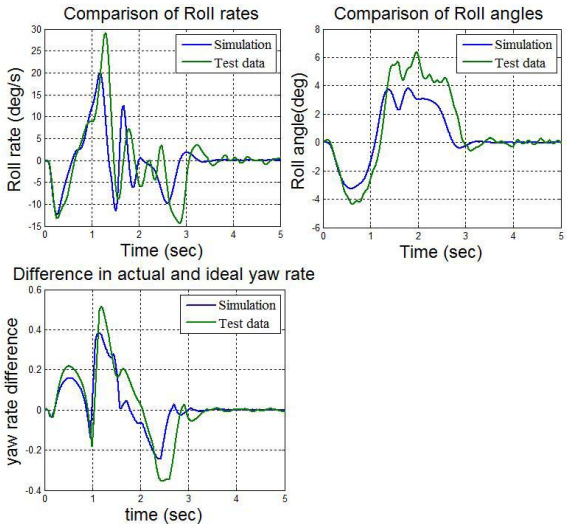


Figure 6. Sine with Dwell -160: Roll Rate, Roll Angle and Vehicle Slip Rate (difference in actual and ideal yaw rates)

Fig. 6 shows that the simulation slightly under-predicts the values of roll rates and roll angles. This could be due to unmodeled phenomena such as tire deformation and bushing stiffness and other compliances not being accounted for in

the model. The plot of vehicle slip rate shows that simulation and test data plots are not identical but they follow the same general trend. The main objective that has to be achieved is that the ESC system model gives performance (in terms of lateral acceleration and yaw rate response) comparable to the system on board the vehicle.

The simulation results showed that though the algorithm used in the ESC model is different from that in the actual on-board system, the vehicle performance, in terms of lateral acceleration and yaw rate response, is comparable.

B. Roll Stability Control Mode

For the YSC mode, test data was available and hence comparisons were made with the ESC system onboard the vehicle.

A new dataset containing the certain car with a high CG was created in CarSim and the ESC roll stability controller was tuned to stabilize this particular vehicle model. The lateral acceleration threshold for the RSC controller was set to 0.7 while the vehicle slip rate threshold was same as that used in YSC (0.18).

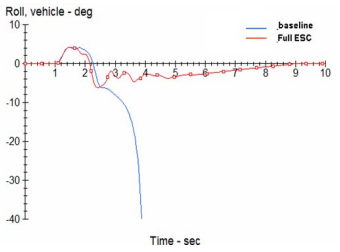


Figure 7. Fishhook (82kmph): Roll Angle

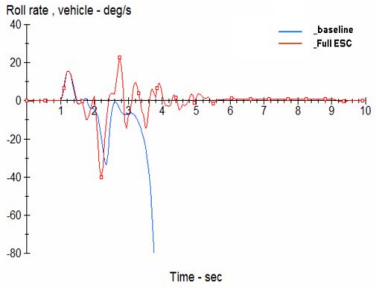


Figure 8. Fishhook (82kmph): Roll Rate

NHTSA's fixed-time fishhook maneuver was used to evaluate the performance of the ESC system in roll stability mode. The certain car with high CG was found to rollover for an initial speed of 82 kmph at the entrance to the Fishhook maneuver. This initial speed was used to compare the response of the baseline vehicle and the vehicle with the ESC system model.

The result shows the comparison plots of various parameters for the vehicle with and without ESC. It is observed that the ESC system prevents rollover by braking the appropriate front wheels. The baseline vehicle rolls over approximately four seconds after the start of the maneuver and hence the data-points for the baseline vehicle are only until four seconds. The roll rate and roll angle increase rapidly as the vehicle goes into rollover.

The Fig. 7 and Fig. 8 show that the ESC system prevents the vehicle from rolling over for an initial speed of 82 kmph in the Fishhook maneuver. To demonstrate that an ESC system increases the stability envelope for a vehicle, The lateral acceleration and yaw rate plots and roll rate and roll angle plots – all come back to zero which indicates that vehicle remains stable and does not roll over.

V. CONCLUSIONS

The main outcome of this study was the development of a simple, functional Electronic Stability Control (ESC) system model for the selected vehicle.

A vehicle dynamics model of a certain vehicle was developed in CarSim. This model was then validated using data from quasi static and dynamic tests. The complete vehicle model was validated using test data from the Sine with Dwell test. There was good correlation between the simulation results and test data.

An ESC system was developed in Matlab-Simulink environment. The simulation results showed that the ESC system model performance was comparable to the actual vehicle, even for very severe maneuvers. Comparison between the vehicle with and without ESC showed that the ESC system improved the roll stability of the vehicle. A simple ABS functionality was also incorporated in the model. The parameters which could be used to tune the model for a different vehicle were listed and the tuning procedure was explained in brief.

REFERENCES

[1] Manning, W.J. and D.A. Crolla, A review of yaw rate and sideslip controllers for passenger vehicles. Transactions of the Institute of Measurement and Control[J], 2007. 29(2): p. 117−135.

[2] Esmailzadeh, E., A. Goodarzi, and G.R. Vossoughi, Optimal yaw moment control law for improved vehicle handling. Mechatronics[J], 2003. 13(7): p. 659-675.

[3] Zheng, S., H. Tang, Z. Han, and Y. Zhang, Controller design for vehicle stability enhancement. Control Engineering Practice[J], 2006. 14(12): p. 1413-1421.

[4] Anwar, S., Generalized predictive control of yaw dynamics of a hybrid brake-bywire equipped vehicle. Mechatronics[J], 2005. 15(9): p. 1089-1108.

[5] Abe, M., Y. Kano, K. Suzuki, Y. Shibahata, and Y. Furukawa, Side-slip control to stabilize vehicle lateral motion by direct yaw moment. JSAE Review[J], 2001. 22(4): p. 413-419.

[6] Ungoren, A.Y., H. Peng, and H.E. Tseng, A study on lateral speed estimation methods. Int. J. of Vehicle Autonomous Systems[J] 2004. 2(1/2): p. 126 - 144.

[7] Au, F.T.K., R.J. Jiang, and Y.K. Cheung, Parameter identification of vehicles moving on continuous bridges. Journal of Sound and Vibration[J], 2004. 269(1-2): p.91-111.

[8] Lingman, P. and B. Schmidtbauer, Road slope and vehicle mass estimation using Kalman filtering. Vehicle System Dynamics[J], 2003. 37: p. 12-23.

[9] Nagai M. The Perspectives of research for enhancing actives safety based on advanced control technology .Vehicle System Dynamics[J], 2007,45(5): 413-431.

[10] Trachtler A. Integrated vehicle dynamics control using active brake, steering and suspension systems[J]. International Journal of Vehicle Design[J], 2004, 36(1): 1-12

A New Improved Algorithm Based on Three-stage Inversion Procedure of Forest Height

Xiang Sun
Department of Space Microwave Remote Sensing System
Institute of Electronics, Chinese Academy of Sciences
Beijing 100190, China
feixiang19913@163.com

HongJun Song
Department of Space Microwave Remote Sensing System
Institute of Electronics, Chinese Academy of Sciences
Beijing 100190, China
hjsong@mail.ie.ac.cn

***Abstract*—The three-stage inversion algorithm and the random volume over ground (RVoG) model can estimate forest parameters, using single baseline polarimetric synthetic aperture radar interferometry (PolInSAR) data. However, the classical method to invert forest height cannot hold the real data, as there are a lot of error data, which is not consistent with practice in theory. Accordingly, the results of classical inversion method are not accuracy and have many noise points. To overcome this problem, the improved algorithm eliminates the singular points while fitting the least square line to determine the location of the ground phase in complex unit circle. This new algorithm is more suitable for practical data.**

Keywords-PolInSAR; forest parameters inversion; topography accuracy;

I. INTRODUCTION

Polarimatric synthetic aperture radar (SAR) interferometry is a combination of polarimetric SAR and interferometric SAR technology [1]，which can extract the three-dimensional (3D) structure information as well as scattering information of the observation targets[8]. POLINSAR is sensitive to scatter location, distribution, and to movement and change characteristics with SAR interferometry, simultaneously is sensitive to scatter structure, direction, symmetry, texture, and dielectric constants characteristics with SAR polarimetry on [2]. PolInSAR is a new earth observation technology that is widely used in estimation parameters of forest. Forest plays an important role as a natural resource in the carbon (biomass) storage and the carbon dynamic cycle. Doing statistics of the total capacity of forest is crucial for environmental and climatic changes. Recently, PolSAR interferometry has shown promise of estimating forest heights based on a random volume over ground model (RVoG), which is introduced by Cloude and Papathanassiou in 2001[3]. RVoG model presumes that two layers combine the vegetation field: vegetation layer and ground layer as shown in Figure 1. The vegetation layer is modeled as a layer of thickness h_V containing a volume of with randomly oriented particles and scattering amplitude per unit volume m_V. The vegetation layer is located over the gound layer positioned at $z=z_0$ with scattering amplitude m_G. In 2003, Cloude and Papathanassiou provide a new geometrical approach for the inversion of RVoG model, widely used for estimation of the height of vegetation using PolInSAR data. However, the three-stage inversion has too many unstable parameters. More over, the result will have noise point. By now, some literatures tend to improve the three-stage inversion algorithm.

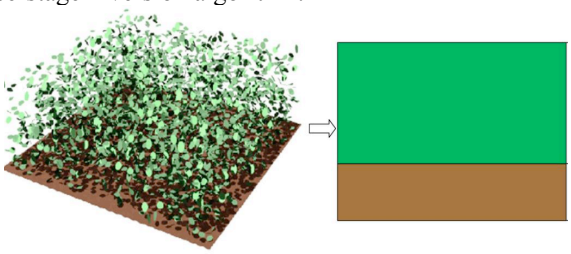


Figure 1. The scattering model of two vegetation layers.

This paper aims to improve the accuracy of three-stage inversion algorithm.

II. THE THREE-STAGE INVERSION ALGORITHM

A. RVoG Model

In order to use PolInSAR data to estimate forest parameters, it is necessary to build a simple and realistic scattering model. The two-layer RVoG model, which combines the vegetation layer and the ground layer, satisfy the demands of inversion model. As is mentioned in introduction, RVoG model has vegetation layer as a layer of thickness h_V, whose scattering amplitude per unit volume is m_V, while the ground layer high h_G with scattering amplitude m_G.

Overlooking the influence of time decorrelation and noise, the complex polarimetric interferometric coherence $\gamma(\omega)$ is given by :

$$\gamma(\omega) = e^{j\phi_0} \frac{\gamma_V + m(\omega)}{1 + m(\omega)} \tag{1}$$

In which the projecting vector that decided by the way of polarimetry is ω, the interferometric phase that correlated with the height of terrain is ϕ_0. And $m(\omega)$ is the ratio of ground to volume scattering. The only parameter that correlates with upper vegetation layer is the complex coherence for the volume γ_V, which is defined as follow:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.111

$$\gamma_V = \frac{I}{I_0} = \frac{\int_0^{h_V} e^{jk_z z'} e^{(2\sigma z')/\cos\theta_0} dz'}{\int_0^{h_V} e^{(2\sigma z')/\cos\theta_0} dz'} \quad (2)$$

The extinction coefficient σ corresponds to a mean extinction value for the vegetation layer. k_Z is the effective vertical interferometric wavenumber after range spectral filtering, which depends on the imaging geometry and the radar wavelength

$$k_z = \frac{4\pi}{\lambda \sin\theta} \delta\theta \quad (3)$$

δθ is the incident angle, which change with the baseline B.

B. The Three-stage Inversion

The classical forest height inversion algorithm is the three-stage inversion, which is proposed by Cloude and Papathanassiou in 2003 [4]. The algorithm divides inversion process into three stages: least squares line fit; vegetation bias removal; height and extinction estimation.

According to the RVoG model, the polarimetric variation changes the place where the interferometric coherence γ(ω) is situated in the complex plane. However, all of the γ(ω) are situated on one line. In that case, we can rewrite γ(ω) as:

$$\gamma(\omega) = e^{j\phi_0}(\gamma_V + L(\omega)(1-\gamma_V)) \quad (4)$$

where the ratio of ground to volume scattering intensity $L(\omega) = m(\omega)/(1+m(\omega))$, varies from 0 to 1. While $L(\omega)=0$, γ(ω) is only related with the volume scattering part. While $m(\omega)=1$, γ(ω) is only related with the ground scattering part. As we can derive from (4), while $L(\omega)$ varies from 0 to 1, the complex interferometric coherence moves from $\exp(j\phi_0)\gamma_V$ to $\exp(j\phi_0)$, as shown in Figure 2.

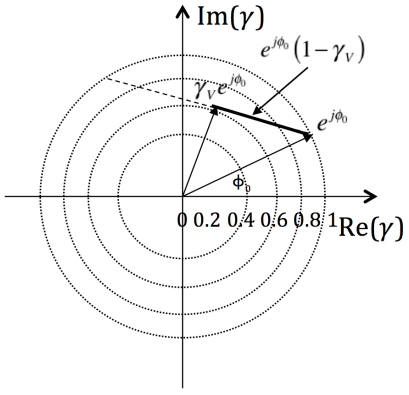


Figure 2. Line model for polarimetric variation of interferometric coherence.

1) Stage 1: Least squares line fit

Using PolInSAR data to estimate the line, we should first know the complex coherence $\gamma_i (i=1,2,\ldots,N)$. According to the least square method we can get both the cross points of the line and the complex unit circle [5].

2) Stage 2: Vegetation bias removal

In this stage, we choose one of the cross points to be the ground scattering phase point. As we know about the characteristic of the scattering system, the amplitude of the ratio of ground to volume scattering is relatively small in HV channel, while it is larger in the HH+VV and HH-VV channels. Consequently, the ground phase point should be further to γ_{HV} rather than γ_{HH+VV} or γ_{HH-VV}. In the ideal situation, γ_V should be the point where $m(\omega)=0$. However, the lack of information of $m(\omega)$ lead to the inaccurate of γ_V. We presume there is no weight of ground scattering in HV channel, then the volume scattering coherence is approximately:

$$\hat{\gamma}_V = \gamma_{HV} \exp(-j\phi_0) \quad (5)$$

According to the RVoG model, the volume scattering coherence $\hat{\gamma}_V$ varies because of the height of forest and the extinction coefficient σ, which is shown in the equation (2). Consequently, we can build a look-up table using the relationship of equation (2).

3) Stage3: Height and extinction estimation

Until stage 2, we can get two volume scattering coherences. One is an estimation value; the other one is in the look-up table. Compare the estimation value of each pixel with the values in the look-up table. Find the closest one, and then we can secure estimates of the height and extinction.

III. THE INVERSION BY IMPROVED ALGORITHM

A. Problems of the Three-stage Inversion

As we can see in Figure 5 (b), there are a lot of noise points after the three-stage inversion algorithm. It is hard to find the problem with all the wrong estimated points. Accordingly we research one of the noise point (22,35) to find out the reason of the noise points' appearance first. Locate the eight of interferometric coherences: γ_{HV}, γ_{HH}, γ_{VV}, γ_{HH+VV}, γ_{HH-VV}, γ_{opt1}, γ_{opt2}, γ_{opt3} in the polar plane, which are shown as * in Figure 3. The red line is the least square line of the eight interferometric coherences. However, one of the points, the pink one is far from the other seven ones. Obviously the lonely point is a wrong one. Because of the differences between the real data and the theory, there are always errors along the process of estimation. If we use all the eight interferometric coherences to get the line, the blue triangle is the ground phase point, which can get from stage 2. Moreover the purple circle is the estimated γ_V from the three-stage inversion algorithm with all the data. The estimation of volume scattering coherence is not near the least square line. It is obviously a wrong one. There are a lot of pixels, whose interferometric coherences are not as promising as the data presumed in the theory. That is the reason why there are a lot of noise points in the result of the classical three-stage inversion algorithm.

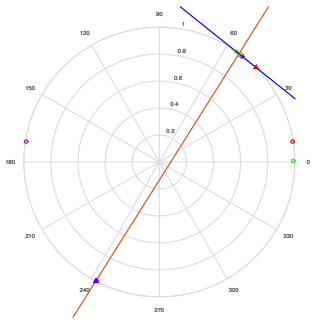


Figure 3. Sample coherence loci for point (22,35).

B. *The Improved Algorithm*

After the analysis above, getting rid of the error of data should optimize the three-stage inversion algorithm. Continually, we use point (22,35) as an example to see how to work. As we can see in Figure 3, the lonely point of the interferometric points is a singular point. It is the error point of the data, so we should overlook it. The blue line in Figure 3 is the least square line of the seven points without the singular point. And the red triangle is the new ground phase point. After the steps of inversion, we can get the volume scattering coherence with out the singular point, which can be seen in Figure 3 as a red circle. It is much closer to its least square line, which mean the step leaving the singular point is effective.

C. *Inversion Steps Using the Improved Algorithm*

The geometrical interpretation of the improved inversion algorithm is shown in Figure 3. The following procedure demonstrates the major steps of the improved inversion:

1) Calculate the interferometric coherence, then move the singular points.

2) Least squares line fit.

3) Vegetation bias removal.

4) Height and extinction estimation.

D. *Problems of New Inversion with Improved Algorithm*

Even with the improvement of getting rid of the singular point, there are still some noise points in the final inversion result. The improved algorithm cannot solve all the problems. In order to smooth the final result, we use a 5×5 boxcar filter.

IV. RESULTS

A. *Result of the Three-stage Inversion Algorithm*

In the practical operation, we choose a set of L-band full polarimetric SAR interferometry data for algorithm validation, which is produced by ESA PolSARpro software. Simulation parameters: platform height 3000m, vertical baseline 1m, horizontal baseline 10m,incident angle 45° , carrier frequency 1.3GHz, forest height 18m,the image size is 105×141. The Pauli base RGB color composition map is shown as Figure 4. The white line in the image is the central line of it.

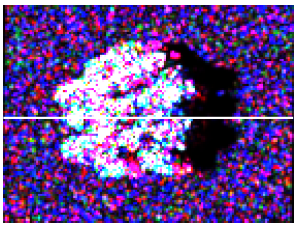


Figure 4. Pauli base RGB color composition map.

The result of three-stage inversion of the selected polarimetric SAR interferometry data is shown in Figure 5 (a).

B. *Result of the Improved Algorithm*

In practical operation, the result of moving the singular points is shown as Figure 5 (b). There are much less noise points in the final inversion image. Figure 5 (c) is the result of 5×5 boxcar filtered. We can see the boxcar filter shows a much specific result without any noise point.

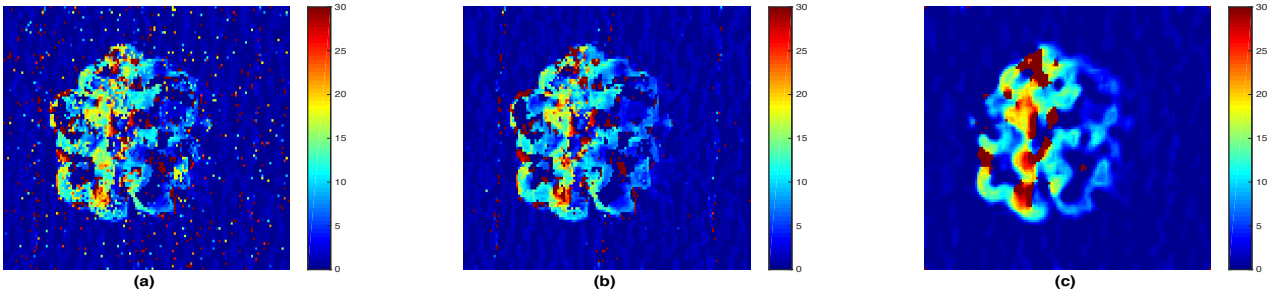


Figure 5. Results of inversion: (a) result of the traditional three-stage inversion (b) result of the improved algorithm after moving the singular points (c) result after the 5×5 boxcar filter.

In order to make the result easier to qualification, using the central line of azimuth direction to analysis the differences between the original three-stage inversion and the improved algorithm. As can be seen from Figure 6, the results of the original and improved algorithm are shown as the red and blue line respectively. Comparing with the improved one, the result of the original algorithm is sharper and has more peaks, which means that the original result has noise points. However after improvement, the result is smoother and does not have noise points.

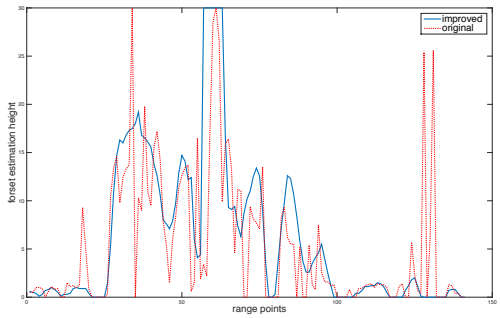


Figure 6. The inversion results of the central line.

V. CONCLUSIONS

The improved inversion algorithm is based on the classical three-stage inversion algorithm and the RVoG model. Previous study has shown the inversion need interferometric coherences as parameters. And the theoretical location of interferometric coherence in complex plane is on a line. However, the classical model and algorithm do not consider the complex situations of real data. The real data have a lot of errors [6][7]. The improved algorithm, which is introduced in this paper, considers the error caused by real data. While fitting the least square line, it is crucial to get rid of the singular data. In this way, the estimated line and the estimated ground phase is much close to the real one. After the improving the original algorithm, the inversion result decrease 2715 noise points which is 18.34% of all the points in the image. Therefore the result the forest height inversion procedure would be much more accuracy and reliable.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61431017) .The authors would like to thank the editors and the anonymous reviewers for their constructive comments that significantly improved the quality of this paper.

REFERENCES

[1] Cloude SR, Papathanassiou KP, “Polarimetric optimisation in radar interferometry”, Electron. Lett., 1997, 33, (13), pp. 1176-1178

[2] Papathanassiou KP, Reigber A, and Scheiber R. “Airborne polarimetric SAR interferometry [C]”. Geoscience and Remote Sensing Symposium Proceedings, Seattle, WA USA, Vol.4: 1901-1903, July 1998.

[3] K. P. Papathanassiou,S. R. Cloude. Single baseline Polarimetric SAR Interferometry. IEEE Transactions on Geoscience and Remote Sensing,2001, 39(11):2352-2363

[4] Cloude SR, Papathanassiou KP. “Three-stage Inversion Process for Polarimetric SAR Interferometry [J]”. IEE Proc on Radar, Sonar and Navigation, vol 150 (3) :125- 134,2003.

[5] Isola, M. and S.R. Cloude. Forest Height Mapping Using Space-Borne Polarimetric SAR Interferometry[C]. in Proc. IGARSS'01. 2001.

[6] Lee S K, Kugler F, Papathanassiou K P, et al.. Quantification of temporal decorrelation effects at L-band for polarimetric SAR interferometry applications[J]. IEEE journal of selectec topics in applied earth observations and remote sensing, 2013, 6(3): 1351-4367.

[7] Thomas Flynn, Mark Tabb, and Richard Carande. Coherence Region Shape Extraction for Vegetation Parameter Estimation in Polarimetric SAR Interferometry. IGASS, 2002:2596-2598.

[8] Wu Yi-rong, Hong Wen, and Wang Yan-ping. The current status and implications of polarimetric SAR interferometry [J]. *Journal of Electronics & Information Technology*, 2007, 9(5): 1258-1262.

The research of feedback - feedforward iterative learning control in hydrodynamic deep drawing process

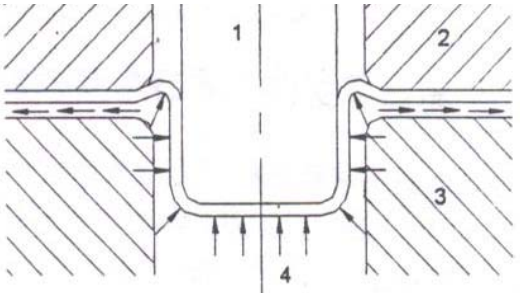
Songwei Shi
Wuxi Traffic Branch Automatic Engineering Department
Jiangsu United Vocational Institute
Wuxi, China
wuxigulf@163.com

***Abstract*—This paper firstly introduces the characteristics of hydrodynamic deep drawing（HDD）, which is an important sheet metal forming technology, then points out the necessity of the chamber pressure control. Secondly considering the characteristics of the drawing action is repeated, iterative learning control (ILC) is the proper algorithm. Then introduces the concept of iterative learning control and feedback - feedforward iterative learning control to solve the delay and improve system robustness. Finally, the computer iterative learning control algorithm implementation process is given and the effectiveness of the algorithm is verified by simulation.**

Keywords- Feedback;feedforward;Iterative Learning Control; hydrodynamic deep drawing; pressure control

I. INTRODUCTION

Hydrodynamic deep drawing process is showed in Figure 1, the dies are filled with liquid, when the punch is moving downward, the liquid in the chamber produces a relatively stress, so the blank can be tightly affixed to the punch in the die to form a favorable friction effect; Also between the dies and the lower surface of the blank sheet a fluid lubrication is produced to reduce the harmful friction, so that the limit of sheet metal forming is greatly improved to obtain high-quality parts, while reducing the defects in traditional drawing process. HDD is suitable for complex shaped parts such as parabola-shaped parts, tapered parts, auto parts and other covering parts requiring high surface quality, and for materials which requires intermediate annealing in multi-pass deep drawing process such as stainless steel.



1-punch 2- Blank holder 3- Dies 4- chamber

Fig.1 The process of hydrodynamic deep drawing

In the literature [1], the author makes finite element simulation in hydrodynamic deep drawing process of cylindrical cup parts and pyramid box shaped part by using a dynamic explicit finite element software package named LS-DYNA3D and the following conclusions are obtained:

In hydrodynamic deep drawing process, the main role of the pressure of the liquid in the chamber is:

(1) In deep drawing process, no dramatic changes of the thickness of sheet metal.

(2) To promote fracture dangerous point transiting from sheet metal at the corner of the punch to the mouth of dies, that dangerous rupture point is moved upward.

(3) In the sheet metal deep drawing process, due to the liquid pressure in the chamber the sheet at the die radius can avoid to touch the die radius, and can reduce the friction and radial sheet tensile stress .

Also pointed out the tubular parts the main reasons of various failure modes in hydrodynamic deep drawing: the tubular member is filled deep drawing process, the breakage of the situation and set the pressure of the liquid filled chamber has a great relationship in order to eliminate the break, not only to control the highest and lowest values of the fluid pressure chamber filling, but also control its dynamic process of change;

Last pointed out: considering the fixed side gap of the parts, The pressure of chamber is provided with a certain range, and the range of increases narrowed drawing ratio of the higher pressure control requirements. Filled chamber pressure plays an important role in filling the drawing process, it is also an important factor in sheet metal forming limit increase. So the chamber pressure is the main control parameter in the hydrodynamic deep drawing process.

In deep drawing process the action of the equipment is repeated and with the same batch of products, the given pressure trajectory of the chamber also remains constant. From the above characteristics, on the system may employ iterative learning control strategy for the main control parameters - chamber pressure control. Features of the control strategy is without knowing the actual model structure and parameters of the system, through the input and output data of the system, the equipment can be well controlled by learning several deep drawing process.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.112

II. ITERATIVE LEARNING CONTROL ALGORITHM

Iterative learning control is developed in the study of robot trajectory control problem, suitable for the controlled object has a certain repetitive motion properties, can be completely tracked on a finite time interval. Iterative learning control is adapted to the controlled object has a certain repetitive motion properties, may be performed on a finite time interval is completely tracked. The control actions on the object amend the undesirable control signal by using the error between the system output and the given trajectory control signal to produce a new control signal so that tracking performance is improved.

The most prominent feature of iterative learning control is possible to use relatively simple iterative learning algorithm to amend the control actions even for a strong nonlinear dynamic system with uncertain parameters. The control system tries several iterations so that the system output can track the target object with high accuracy. Which is essentially belongs to no model control (model-free) method, that is, only using the system input and output for the controller design and the controller structure does not depend on the controlled object, so it can be essentially applied to nonlinear control systems.

Robustness of the practical application of iterative learning control should be guaranteed, so the robust convergence of practical iterative learning law is needed. As previously mentioned, the conclusion of iterative learning control convergence is obtained under the premise of many assumptions. But the actual process has the presence of measurement noise and uncertainties disturbance, the initial positioning system is difficult to be completely accurate. Without the interference the iterative learning law can converge but it may not converge and even diverge in the interference environment.

In any system in the uncertainty system family, the iterative output trajectory of an iterative learning control system can still converge to the area of the desired trajectory even with the various disturbances and initial state errors. When the disturbances have been gradually eliminated or repeated, the input trajectory can still converge to the desired trajectory as close as possible, it is called the iterative learning control design has the Robustness to anti input and output disturbances and initial offset. [2]

The description of Feedback - Feedforward iterative learning control which is one kind of the robust iterative learning control algorithms, the simulation and analysis of the actual system using this algorithm are given below.

III. FEEDBACK - FEEDFORWARD ITERATIVE LEARNING CONTROL ALGORITHM

For some saturated actuator systems, high-gain feedback will cause excessive control signal, due to the limiting action actuator saturation, the learning speed of the control system will be affected, so the high gain feedback is meaningless. At this time, if a feedback loop is put into open-loop iterative learning control system, forming a feedback - feedforward iterative learning control system is an effective way to improve the tracking performance.

There are two feedback - feedforward iterative learning control algorithm, the first algorithm is described as follows: Set feedback controller:

$$u_{fb},k(t) = h_{fb}(e_k(t)) \tag{1}$$

Feedforward controller uses an open-loop learning law:

$$u_{ff,k+1}(t) = u_{ff,k}(t) + h_{ff}(e_k(t)) \tag{2}$$

In this case, the system input can be written

$$\begin{aligned} u_{k+1}(t) &= u_{ff,k+1}(t) + y_{fb,k+1}(t) \\ &= u_{ff,k}(t) + h_{ff}(e_k(t)) + h_{fb}(e_{k+1}(t)) \end{aligned} \tag{3}$$

Another use feedback - feedforward iterative learning control algorithm is as follows:

Feedforward controller uses an open-loop learning law

$$u_{ff,k+1}(t) = u_k(t) + h_{ff}(e_k(t)) \tag{4}$$

At this time, the system input is

$$u_{k+1}(t) = u_k(t) + h_{ff}(e_k(t)) + h_{fb}(e_{k+1}(t)) \tag{5}$$

IV. THE ALGORITHM USING IN THE DEEP DRAWING PROCESS

Iterative learning algorithm can be easily implemented on a digital computer. The deep hydrodynamic control equipment has been equipped with a programmable controller (PLC), which like a digital computer, thus iterative learning algorithm can be programmed by the PLC.

V. SIMULATION AND ANALYSIS OF THE ALGORITHM IN MATLAB

The controller of feedback-feedforward iterative learning control system consists two parts, one is feedback controller another is feedforward controller, which uses the open and closed loop with the learning law. Feedback controller is used to implement the system a calming effect, it allows the system output does not deviate too far from the desired trajectory. Under the sedation of the feedback controller, the feedforward controller can be expected to quickly achieve full track tasks.

By using graphical simulation environment integrated with the Simulink in software MATLAB, a dynamic model of hydraulic deep drawing is established. Then a variety of iterative learning control algorithm is simulated for the model and the results are analyzed, which verify the validity of the iterative learning control hydrodynamic deep drawing process pressure control.

Based on the simulation models are given in the literature [3, 4, and 5] and the experimental data of hydrodynamic deep drawing device, the final simulation model is established below in figure 2.

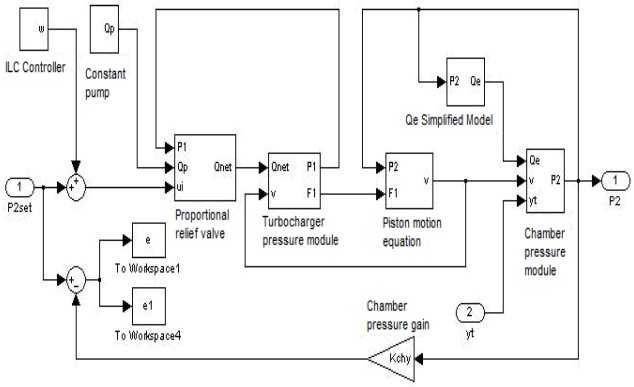
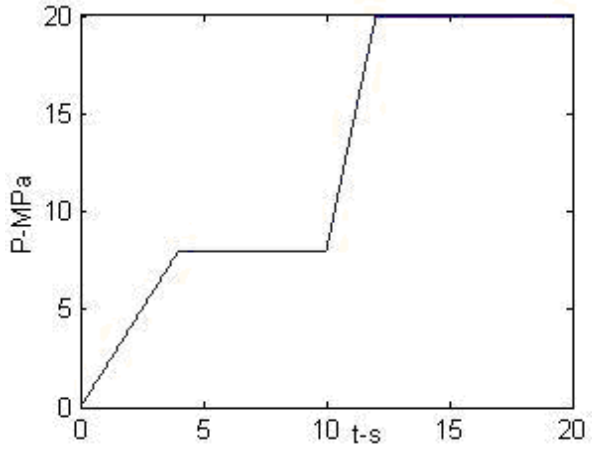
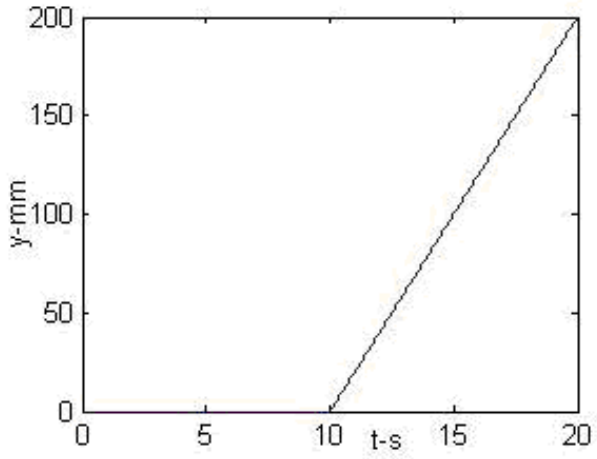


Fig.2 System simulation model

Before the start of the simulation, two given signals are determined: The figure 3(a) is the given signal of the chamber pressure and (b) is the given signal of the punch movement.

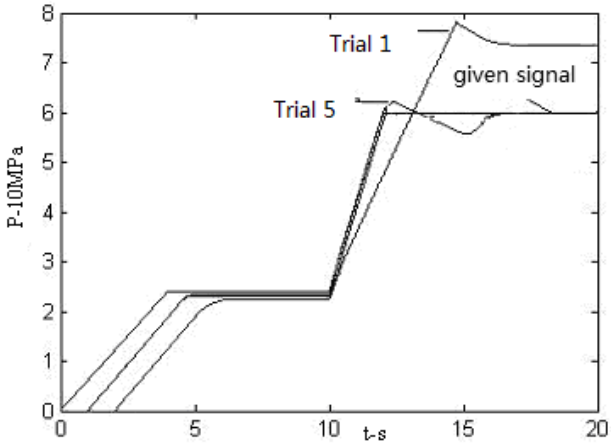


(a) The chamber pressure

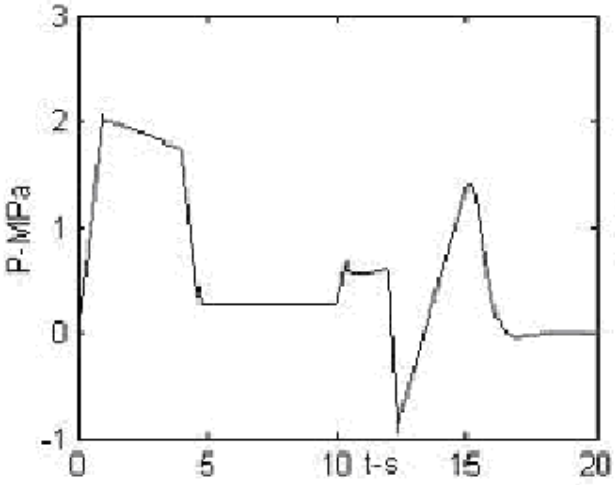


(b) The punch movement

Fig.3 The given signals using in simulation



(a) Press-time

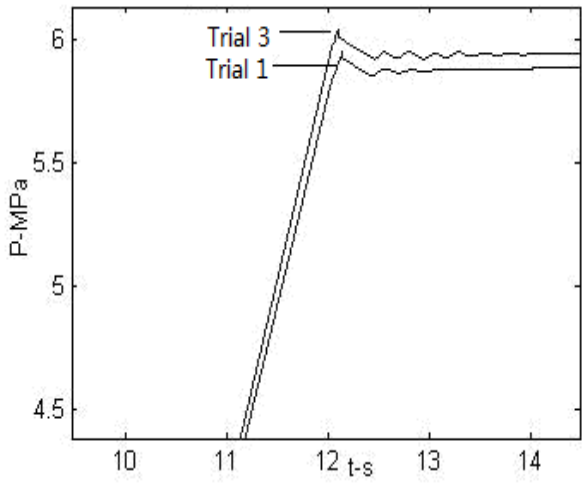


(b) Pressure error-time (Trial 5)

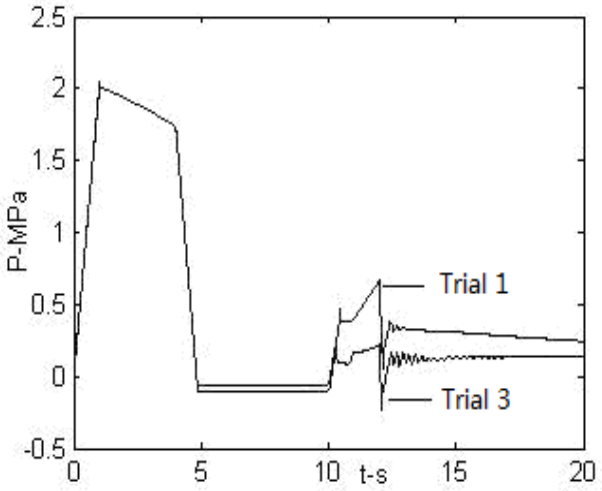
Fig.4 Result of classical PD iterative learning control

The figure 4 is the simulation results of using the PD controllers designed by the classical feedback control theory to achieve the system pressure closed-loop control in feedback loop without any controller in feedforward loop. In the PD controller K_P=1, K_D=0.4.

The figure 5 is the simulation results of using PID controllers designed by the classical feedback control theory to achieve the system pressure closed-loop control in feedback loop and P iterative learning controller is used in feedforward loop. In the PID controller P=20, I=1, D=0. In the P controller the learning coefficient Kp=1.



(a) Press-time



(b) Pressure error-time

Fig.5 Result of feedback-feedforward iterative learning control

With the Comparison of the ordinary PD iterative learning control and this new hybrid control strategy, the results show it really can improve tracking performance while accelerating the learning speed and improving the robustness of the system.

VI. CONCLUSION

This paper introduces the basic concepts of hydrodynamic deep drawing, and points out the importance of chamber pressure control. Then the conception of iterative learning control is given and feedback - feedforward iterative learning control algorithms is selected to solve practical system hysteretic effect on the control performance and improve the robustness of the system simultaneously. Finally, By using Matlab, the system model is simulated to verify the feedback - feedforward iterative learning control algorithm is effective.

REFERENCES

[1] Li Hui Lang.filled in numerical simulation of material forming process of drawing.Harbin Institute of Technology, doctoral thesis,1998.

[2] Shao Cheng, Gao Fu-Rong, Yang Yi. Robust Stability of Optimal Iterative Learning Control and Application to Injection Molding Machine. Automation Journal. 2003, vol. 29, pp. 72-79.

[3] Danian Zheng, Heather Havlicsek, Andrew Alleyne. Nonlinear Adaptive Learning For Electrohydraulic Control Sytems, Mechatronics, IEEE/ASME Transactions , 1999, Vol. 4 , pp.316-317.

[4] Heather Havlicsek, Andrew Alleyne. Nonlinear Control of an Electrohydraulic Injection Molding Machine via Iterative Learning. Proceedings of the American Control Conference,San Diego, California,1999, pp. 176-181.

[5] Danian Zheng Andrew Alleyne Learning Control of an Electro-hydraulic Injection Molding Machine with Smoothed Fill-to Pack Transition Proceedings of the American Control Conference,Chicago,Illinois,2000, pp.2558-2562.

Electrical Servo Screwdown Control System on Cold Rolling Mill for Traveler Substrate

Xueyang Yu, Jing Hui
Key Laboratory of Advanced Process Control for Light Industry,Ministry of Education
Jiangnan University,Wuxi,China
766084033@qq.com ; jingh@126.com

***Abstract*—A novel roll position control method, based on electrical screwdown, is proposed to solve the problems such as the complicated operation and poor positioning precision of the manual screwdown on the cold rolling mill for traveler substrate. The new system adopts PLC as the main controller,servo driving system as the actuator, and realizes the flexible adjustment of the roll gap by the proper programming of PLC and the touch screen.The construction and operation principle of position control system, the hardware realization and the software design are detailedly introduced.Through engineering practice proves that the system has 0.001mm of the high control accuracy, suiting to requirements of the thickness of 0.4mm-2mm of the traveler substrate.**

Keywords—traveler substrate;cold rolling mill; electrical servo screwdown; position control

I. INTRODUCTION

Traveler is the main equipment of twisting and winding, it has an important impact on the yarn quality[1].Substrate for cold rolling mill forming process is rolling the round wire which has been drawn to a certain diameter into a specific type substrate.This step is significant in the entire production process of the traveler product.The product quality and appearance level are directly affected by the rolling forming precision , the setting of roll gap is the premise to ensure the traveler substrate thickness precision. Therefore, roll position control become an important part of the traveler production.

In the traditional production process ,the roll gap were set by the operator who manually adjusted the screwdown [2].The precision of setting is just rely on the experience of the operator, and for different sizes of traveler must be through debugging for many times. Obviously, manual adjusting the screwdown exists some problems such as the complicated operation and low regulation accuracy.

To achieve a high-precision traveler substrate, a novel roll position control system is introduced in this paper. The system adopts servo motor to drive the screwdown. Through compiling the reasonable position control algorithm to realize the automatic control of the roll position. The new system positioning process is simple, high control precision ,and laying the foundation for the automation control in the entire production line.

II. CONTROL SYSTEM COMPONENTS AND OPERATING PRINCIPLE

Schematic diagram of roll position control system is showed in Figure 1.The system adopts PLC as the main controller,servo motor as the actuator, rotary encoder as the detecter, the F940GOT touch screen as the human-machine interface (HMI), and realizes the information exchange by the FX_{2N}-485-DB communication expansion module.Giving the position signal by the HMI while the substrate thickness is determined. Then PLC calculates and outputs the pulse signal to the servo drive, servo motor according to the pulse amount and pulse frequency to regulates the screwdown, the PG feed back the current position signal to the PLC, PLC changes its pulse output through the position control algorithm,therefore changes the servo motor speed.

In the article, regarding the rolling baseline as the reference line.Precisely control the gap bewteen the two rolls by controlling the distance of the upper roll relative to the baseline.According to the needs of the system,roll position can only be adjusted in a certain range, the start and end point of the upper roll must be set properly. To avoid the servo system send the wrong positioning completion signal,three consecutive positioning signals should be received in the positioning process[3].

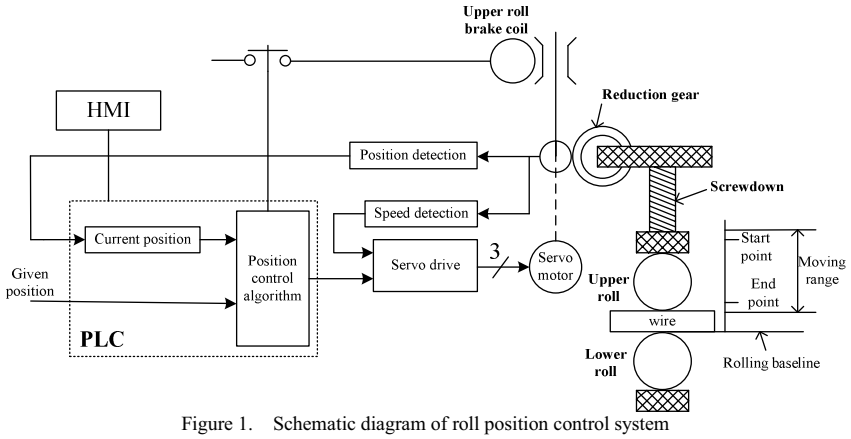
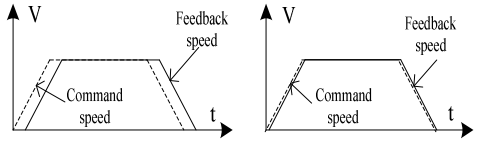


Figure 1. Schematic diagram of roll position control system

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.113

A. Linear tracking of roll position control

The controlled object is automatically controlled to the pre-given target position, the deviation of the actual position and target position is maintained within the tolerance range, this control process is called automatic position control (APC)[4]. In the cold rolling mill system,roll position control is mainly achieved by precisely control the motor speed.Motor speed control is realized by the trapezoidal velocity chart. As shown in Figure 2.



Without speed feed-forward link With speed feed-forward link

Figure 2. Comparison of the speed curve

Ideally,the feedback speed is completely tracking the command speed, the position deviation is zero. However, due to the transmission device and other factors affect the response speed,the feedback speed must lag behind the command speed,which will result in the increases in the position deviation.Thus, adding the speed feed-forward correction link to the roll position control, so we can adjust the speed feed-forward coefficient to reduce the tracking deviation. Figure 2 is the comparison of velocity curve.

B. Ideal model of roll position control

In the roll position control system, the requirements of the ideal position is completing the positioning action in a short time.

Figure 3 is the ideal positioning process chart. Position deviation is S, the initial position deviation is S_0, the maximum acceleration of motor is a_m, the maximum speed is v_m.To make the screwdown reach the given position as soon as possible,so the screwdown should be at maximum speed as many as possible.To guarantee the adjustment time, the motor should be executed with maximum acceleration and deceleration,so that the screwdown can happen to stop at the specified position.

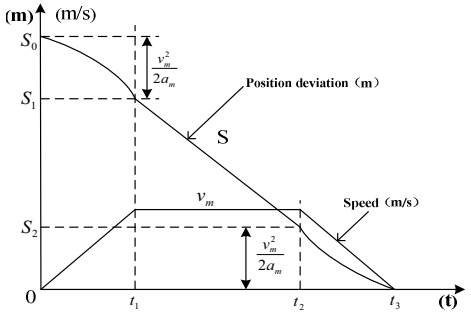


Figure 3. Ideal of positioning process

Ideally, the speed in the acceleration phase is:

$$v = a_m t \tag{1}$$

$$S = S_0 - \int_0 v dt = S_0 - \frac{1}{2} a_m t^2 \tag{2}$$

At this point, the actual acceleration time is:

$$t_1 = \frac{v_m}{a_m} \tag{3}$$

When the servo motor reaches the maximum speed, the position deviation is:

$$S_1 = S_0 - \frac{v_m^2}{2a_m} \tag{4}$$

If the screwdown does not reach the given position at this time, then the motor should continue to move at v_m .To ensure the high control precision of the system, it is important to choose the starting deceleration time of the motor. In the system, the maximum acceleration is equal to the maximum deceleration, so the acceleration distance is the same as the deceleration distance, the deceleration distance is:

$$S_2 = \frac{v_m^2}{2a_m} \tag{5}$$

Ideally, the deceleration speed is:

$$v = \sqrt{2a_m S} \tag{6}$$

The system begin to decelarate in S_2 at the maximum deceleration, and just completing the positioning process while the motor speed is zero.But in practical engineering applications,the effects of sampling and transmission device and other factors will make the switching time could not in the ideal deceleration point, and it may extend the positioning time and decrease the positioning precision[5].

By equation(6) is known, $v = f(S)$ is a parabola curve when S between S_2 and 0. It is difficult to control the deceleration process in the allowable deviation range while the position deviation is very small. So we can set the appropriate speed curve method to solve this problem.

C. New position control algorithm

The requirements of position control: high accuracy and no overshoot, so the system can adopts the method of "block deceleration control" to realize this requirement. In the position control system, the APC algorithm can be simply expressed as:

$$v = f(S) \tag{7}$$

Above, v is output speed of the control algorithm , S is the position deviation.

The basic idea of the new position control algorithm is that in the vicinity of $S = 0$, making $\mathrm{d}v/\mathrm{d}S$ equal to a constant k. The speed ratio would not be too large when the screwdown close to the given position (S≈0). Even if the deceleration switching time is later,the system still can able to enter the deceleration range.

When k is determined, if the deceleration of beginning to use $v = \mathrm{k}S$, the actual deceleration may exceeds the maximum deceleration of the motor, which can not be achieved in the system. Thus,it must firstly uses the maximum deceleration a_m to decrease the speed to meet the following condition(8), then switched to $v = \mathrm{k}S$ (As

shown in Figure 4).

$$v = v_2 = \frac{a_m}{k} \tag{8}$$

The system deceleration is exactly equal to the motor maximum deceleration at this time,and the deceleration speed can be switched to the second deceleration phase $v = kS$. Position deviation can expressed as (9) :

$$S = S_2 = \frac{v_2}{k} = \frac{a_m}{k^2} \tag{9}$$

Roll moving distance in the first deceleration phase is:

$$S_3 - S_2 = \frac{1}{2}(\frac{v_m^2}{a_m} - \frac{a_m}{k^2}) \tag{10}$$

The condition of roll switched to the first deceleration phase is:

$$S = S_3 = \frac{1}{2}(\frac{v_m^2}{a_m} + \frac{a_m}{k^2}) \tag{11}$$

Based on the above analysis, the deceleration speed curve of the position control system are setting as shown in Figure 4,and S_1 is the positioning precision of the system requirements.

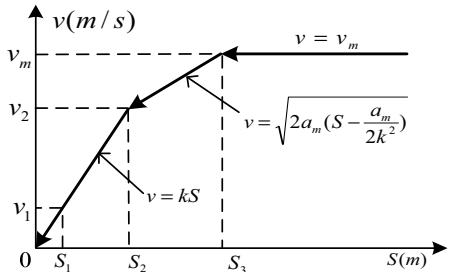


Figure.4 Speed curve of the deceleration phase

III. CONTROL SYSTEM IMPLEMENTATION

A. Hardware Design

PLC controller is the core of the system, it has been widely used in the modern industrial control due to its advantages such as fast response speed, anti-interference ability and simple programming [6].

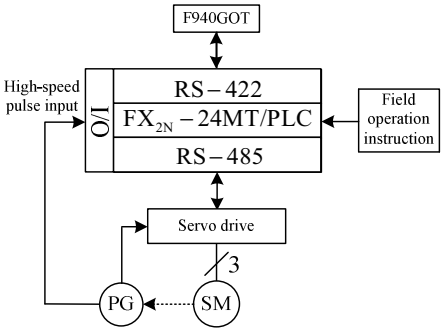


Figure.5 Electrical control system diagram

PLC accepts the field operator instruction,all kinds of feedback status signals and the real-time parameter modification of the HMI. Using servo motor as the actuator, PLC sends the control signals to the servo drive after internal calculation through RS-485.Meanwhile,the servo drive feedback the operation signal to the PLC. Electrical control system diagram is showed in Figure 5.

By calculation, the actual I/O port amount is not more than 32 points, so the system selects the Mitsubishi series FX_{2N}-24MT Type PLC. FX_{2N}-24MT Type PLC belongs to transistor outputs type, which has the general logic control and computing functions, but also has high-speed pulse input, PLSY high-speed cluster direct output (Y0, Y1 port),Ramp variable-frequency soft start / soft brake and other special processing function.

As shown is the definition of the PLC inports in Table I, the system need to occupy 10 inports.PLC input distribution diagram shown in Figure 6.

Table I. Definition of PLC input

Type	Function
Switch	Manual/Automatic
Button	Run, Stop, Jog
Fault signal	Servo fault
Control signal	Pulse input,Servo-ready,Positioning completion Start、End point limit

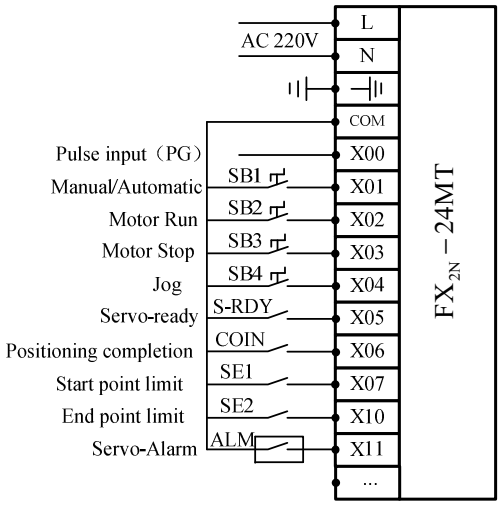


Figure.6 Distribution diagram PLC input

Screwdown is drived by the high-precision AC servo motor, the system selects Yaskawa SGDM-75ADA servo drive and the corresponding motor as the actuator agency. Servo drive is set to the "position control" mode, pulse input mode uses "SIGN + PLUS" pulse signal form. Adding the command tracking control and feed-forward control in the control mode,it can improves the response speed and reduces the position deviation at the most extent.

The system uses Mitsubishi F940GOT touch screen as the HMI, using the corresponding FX-PCS-DU/WIN-C configuration software to design the main screen, parameters setting, status information, alarm information and help screen. The roll gap parameter can be easily modified by the PLC through the communication with RS-422.When the product specification changes, we can directly set up the gap parameter without adjusting any mechanical structure,it greatly improves the production efficiency.

B. Software Design

System control software consists of three parts: PLC control program, touch screen—PLC communications program and touch screen configuration software. Due to the limited space,this article only gives PLC control system flow chart. According to the new control algorithm and roll position control system, automatic position control software flow chat is showed in Figure 7.

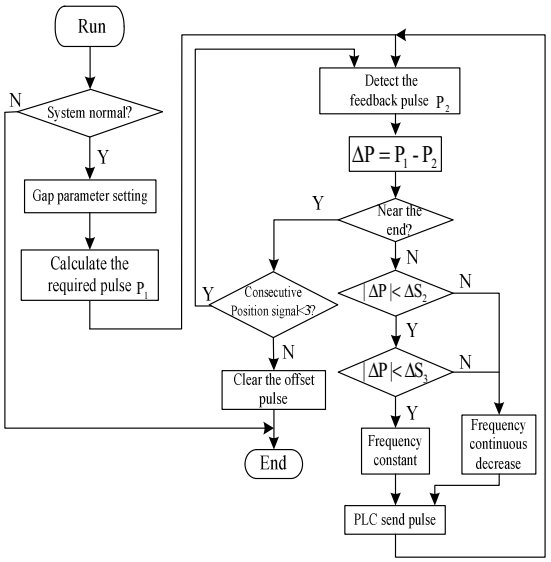


Figure.7 Flow chat of roll position control

After power-on,PLC initializes the system operation state and other variables at first, and then enters the main loop. In the condition of automatic operation, the system is detected whether it is in the ready state .If it is, PLC will calculate the required amount of pulses according to the roll gap which is set in the touch screen,and read out the pulses which has been in operation through the PG,then PLC computes the remaining number of pulses of the system .If the screwdown doesn't in the vicinity of the positioning range, according to the remaining number of pulses to determine whether the system is in a constant speed phase, first deceleration phase or the second deceleration phase, PLC will send the corresponding pulse depend on the corresponding phase. When the screwdown in the vicinity of the positioning range, to ensure the positioning precision of system, three consecutive positioning signals must be received to judge whether the positioning process is completed , then system clears the offset pulse, the program stops running.

IV. CONCLUSION

Aiming at the requirements of the roll gap setting of the cold rolling mill,a new type of roll position control system is designed in this peper.The system changes the traditional model that manually adjust the screwdown to set the roll gap,adopts high-performance PLC as the core controller to improve the system reliability in the complex industrial environment by compiling reasonable software program; Uses high-precision positioning servo motor,to further improve the accuracy of position control; The flexible adjustment of the roll gap is realized by the parameters setting of the touch screen.The engineering practice shows that the control effect of the position control system is very good, its control precision can reach to 0.001mm, meeting the requirements of thickness of 0.4mm-2mm of the traveler substrate.

REFERENCES

[1] Cui Shitao,Zhao Panpan.Traveler Analysis of the Influence on Yarn Quality[J].Cotton Textile Technology,2011,39(1):16-18.

[2] Chen Haibo.Research on Automatic Control of the Screwdown System for the Rolling Mill[D].Jiangxi:Taiyuan University of science & Technology.2013.

[3] Yang Kun.Research and Design of Seamless Steel Tube Control System Based on Profield-Bus[D].Wuxi:Jiangnan Universuty.2011.

[4] Jiang Qiong,Zeng Pan.Automatic Control System of the Precise Electrical Screwdown[J].Metallurgical Automation,2014,S2:362-364.

[5] Xu Xingyuan.Research of Position Control System of Rolling Mill[D].Henan:Zhengzhou University.2005.

[6] Han Yuqi.Electrical Control and PLC Application Technology[M]. Nanjing:Southeast University Press,2009.

Research on Tension Control for Coating Line of Optical Films in Dynamic Process

CHEN Ya-wei
Department of Electrical Engineering
Jiangnan University
Wuxi,China
e-mail:18206181389@163.com

HUI Jing
Department of Electrical Engineering
Jiangnan University
Wuxi,China
e-mail: jingh@126.com

Abstract：According to the coating line of optical films at constant speed operation, the tension is stable. However, the tension would appear larger fluctuation in the process of deceleration .On basis of analyzing the relationship among accelerating speed,torque and tension in the coating line , the rolling system dynamics model is set up. Therefore in the process of rolling from constant speed to accelerate, it need to increase the dynamic torque compensation, and the accuracy of dynamic torque compensation determines the stability of the tension. Based on the dynamic torque compensation and real-time measurement of inertia, a real-time dynamic torque compensation control strategy is proposed based on tension.The practical results show that the proposed control strategy that make the coating line rolling in the process of accelerated tension fluctuation is suppressed, and verify the effectiveness of the proposed method.

Key words:optical films, coating line, unwinding tension, dynamic torque compensation, inertia

I. INTRODUCTION

The optical films is optical basilemma that is handled with manufacturing process of coating,drying,and laminated to improve performance of high light transmittance, low degree of fog, high brightness, etc. It is generally used in glasses, mobile phones, computers, LCD TV, etc and make their performance better. There are many kinds of optical film,such as reflecting film, antireflection film, filter membrane,optical protection layer, polarizing film, spectroscopic film and phase film.In order to guarantee the quality of the optical film, we need high quality coating equipment.However,to manufacture coating equipment, manufacturers not only needs strong mechanical design capability, at the same time design high precision of tension control system to meet the requirements.The materials of optical basilemma is generally PET film, but the thickness of PET film is usually a micron level.In order to guarantee the stability of the thin film coating line, tension should be kept in highly accurate and stable micro tension state.

II. PROBLEM FORMULATION

The unwinding system of coating line with two nip rollers studied in this work is equipped with two AC servomotors which are controlled with high performance frequency converters. The considered roller system is part of the processing section of a cardboard packaging machine, Fig. 1. The unwinding system can be divided into three sections: unwind roller, dancer roller , processing roller.

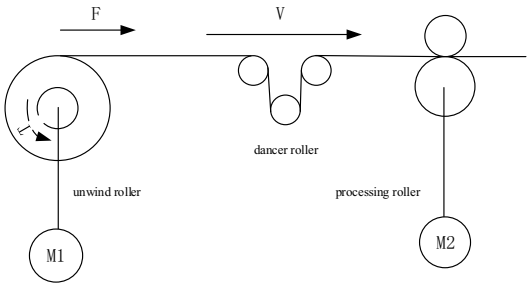


Fig.1Unwinding tension control diagram

In the running process, the unwinding roller with the variable frequency motor 1, the processing roller rotates with the variable frequency motor 2. The tension between unwind roller and processing roller is measured by dancer roller, and controlled by unwind roller in order to keep the tension within the wire constant. In the above described process, the processing roller is left uncontrolled, and works at a constant speed.

A. Mathematical Modeling

When we analyze tension control for coating line of optical films in dynamic process,we can find influence factors include tension torque, friction torque,dynamic torque and motor torque. The mechanical model of unwinding tension is calculated as the sum of various torque if it exists in (1),where M_F represents the tension torque, M_f the friction torque, M_i the dynamic torque, M the motor torque.

$$M_F + M_f + M_i + M = 0 \qquad (1)$$

To obtain the tension torque M_F, the following condition should be met:

$$M_F = F \times R \qquad (2)$$

F represents the optical films tension, and R is the initial volume radius which is measured by photoelectric sensor or ultrasonic sensors and is setted by operator.

M_F which is generated by mechanical friction is not fixed value and changes the value all the time according to the speed value.

M_i is the dynamic torque when the coating line begin to accelerate.The feedforward compensator is used to improve the tracking performance so that no large change in tension during the unwinding dynamic process

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.114

will occur. In this paper, the feedforward compensator for torque is described by:

$$M_i = J \frac{d\omega}{dt} \tag{3}$$

The relationship between the angular velocity ω and speed n is described by:

$$\omega = \frac{2\pi n}{60i} \tag{4}$$

Based on equation (3) and (4), M_i is described by:

$$M_i = J \frac{2\pi}{60i} \frac{dn}{dt} \tag{5}$$

J is rotational inertia, but we can't directly access data. The compensator plays an important role in an intermittent.

B. Calculating rotational inertia

Rotational inertia in coating line is difficult to model accurately, mainly because the system dynamics are influenced by many process variables that vary over a wide range. It is difficult to derive a precise mathematical model for the dynamics of coating line system such that it includes all relevant factors, such as variations of the physical characteristics of the web, friction, slip condition and other non-linear operating variables. Rotational inertia is composed of two parts. One is the inherent inertia J_F, including the rotational inertia of the motor, reducer and drum which is equivalent to the motor shaft parts.The other is variational inertia J_V which is the unwinding roller.Therefore, The total inertia be expressed as

$$J = J_V + J_F \tag{6}$$

The drive torque responses are described with one time constant systems. This is of course not very accurate as the capability of responding to torque demands depends greatly on the voltage reserve available at the very moment.

$$M = M_i + M_f \tag{7}$$

Then,we get the equation to get J_F:

$$\begin{aligned} M_i &= M - M_f = J_F \cdot \frac{d\omega}{dt} \\ &= J_F \cdot \frac{d(2\pi n)}{dt} = 2\pi \cdot J_F \cdot \frac{dn}{dt} \end{aligned} \tag{8}$$

$$J_F = \frac{(M - M_f)}{2\pi} \frac{dt}{dn} \tag{9}$$

Because the unwinding roller is rigid body, the variational inertia J_V can be calculated accurately.

$$\begin{aligned} J_V &= \frac{1}{2} m(R_1^2 + R_2^2) \\ &= \frac{\pi \rho s L(R_1^4 - R_2^4)}{2} \end{aligned} \tag{10}$$

In order to calculate the value Conveniently

$$\begin{aligned} J_V &= \frac{1}{2} m(R_1^2 + R_2^2) \\ &= \frac{\pi \rho s L(R_1^4 - R_2^4)}{2i^2} \end{aligned} \tag{11}$$

Therefore, based on equation (9) and (11), J is described by:

$$J = \frac{\pi \rho s L(R_1^4 - R_2^4)}{2i^2} + \frac{(M - M_f)}{2\pi} \frac{dt}{dn} \tag{12}$$

B.control strategy

The structure of unwinding intelligent tension control (the varistructural intelligent control based on the expert rule) system has shown in Fig. 2. The switch S1 in Fig. 2 is the intelligent tension controller, and can also send out the dynamic speed instruction in "S shape" pattern to the feed speed system.When the switch works on the section 1, the system works on three closed-loop status to keep the tension stable. Moreover when change in speed instruction takes place, it can automatically work in the section 2 to make switch S1 for the dynamic tension control system. According to the tension signals given by Switch S1 (an output of one pulse for each rotation) on unwinding axis and the photoelectric encoder PG on

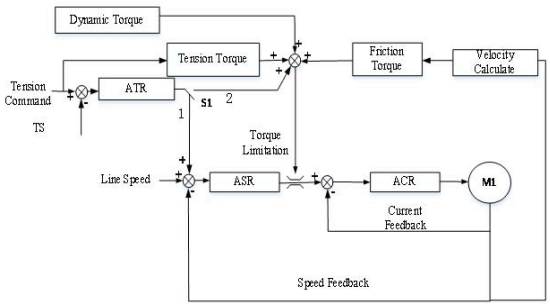


Fig.2 Tension control with dynamic equalizing

feed axis can also perform operation and acquire the value of radius R, to further make corrections on the compensation weights for parameters and feed-forward in the tension control algorithm. In addition, "TS" is the tension sensor; MB and BD are the magnetic powder brake and its fast response PWM driver, respectively; V and F are the strip running speed (m/min) and the tension (N) respectively; n and T are respectively the rotational speed of unwinding axis (rpm) and the brake torque of magnetic powder brake (Nm). The basic operating principle of the system shown in Fig. 2 indicates that when the radius diminishes, the actual tension will get large, and the U is supposed to reduce, so is the corresponding Ub and the exciting current of magnetic powder brake, while T and the exciting current are in an approximately direct ratio relationship, and F=T/R, to ultimately maintain the tension to be equal to the set value and remain invariable.

III. THE STRUCTURE OF THE CONTROL SYSTEM

At present, the control system that PLC controller is the core and upper computer is the real-time monitor has

become a development direction of the industrial automatic control system. Through the communication between PLC and VFD, the system can realize the aim that the technical flow chart, the dynamic data menu and the report diagram can be provided to the upper computer, so that the P L C control system has a good man-machine interface. Through the upper computer reads and writes the PLC data, the system realizes that the on-site data can be collected and transmitted. The process can be controlled automatic and informational, so the prospect of its application is very broad.

Profibus-DP is an international and open standards-based field bus, allowing the data be transmitted quickly and in real-time in the site. It supports the data be transmitted not only in the RS485 wire, but also the optical fiber. The transfer rate can be chosen between 9.6 Kb / s Potter and 12 Mb / s.

Considering the scale of the production line, the configuration of the control syetem is chosen as follows:

- PLC uses the siemens S7-300,CPU module uses the 314C-2DP, which has a Profibus-DP interface itself, Power supply module uses the PS307, AC220V/DC24V, 5A. The S7-300 is the core of the whole control system, receiving all the signals of the sensors, giving the transducers the output signals to control the speed of the motors, providing the datum to the touch screen, and processing the input signals of the touch screen.
- The upper computers use two Siemens touch screens. Because the scope of the production line is long, one is used to control the entire production line, one is used to control the film winding device separately.
- Transducer: the transducers use the Siemens G120.According to the given speed of the PLC and the feedback value of the encoder, the transducer controls the speed of the motor in real time through its own PID regulator, making the whole production line coordinate and synchronous.
- Profibus-DP fieldbus: it connects the PLC with the touch screen, the PLC with the transducer and the pressure sensors, etc.
- The software of programming and communication is Step-7, mainly used to compile the program in the PLC, The software of configuration is WinCC, used to compile the interface of operation.

The overall framework of the control system is shown in Fig.3.

Fig.3 Experimental platform

The parameters selected for the variable frequency drive are presented in Table I.

TABLE I Main Parameter of G120 Inverter

PERF	*Default*	*Function*	*Brief Description*
P0003	3	User Access Level	Expert:For expert use only
P0701	1	Function of Digital Input 1	ON/OFF1
P0702	4	Function of Digital Input 2	OFF3 - quick ramp-down
P1300	21	Control Mode	Vector control with sensor
P1501	R722.3	Change to Torque Control	Selects command source from DI3
P1503	R755.0	Torque Setpoint	
P1520		Upper Torque Limit	
P1521		Lower Torque Limit	

IV. EXPERIMENTAL RESULTS

In the operation of the system,we set the tension of coating line 50N and accelerate the velocity two times. Firstly, we accelerate the velocity from 0m/min to 10m/min, and acceleration is 0.5m/s^2. Secondly, we accelerate the velocity from 10m/min to 100m/min, and acceleration is 1m/s^2. The velocity setting is presented as Fig.4.

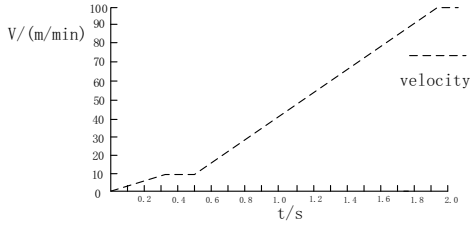


Fig.4 velocity setting

Then, we can observe the fluctuation of tension when we change the velocity. In case of close-loop tension control without feedforward torque compensation, the torque remains almost constant even in acceleration resulting in tension fluctuation in 0.4N at that moment in Fig.5. In Fig. 6 ,the tension overshoot that is changing in 0.2N is effectively suppressed by the compensation of torque command during speed acceleration in system. The comparison of the typical control result of tension control is presented as Fig.5.

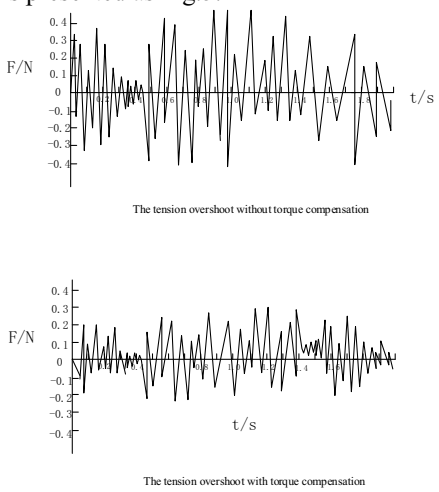


Fig.4 Comparison of the typical control result of tension control

V. CONCLUSION

A new tension control algorithm with tension observer is proposed using observed tension as a regulator feedback. The tension observer is based on the torque balance of a roller stand including the acceleration torque. Using this estimated tension, new tension controller can be constructed with faster dynamic response in case of line speed acceleration or deceleration. The performance of proposed controller is compared with those of conventional open-loop and closed- loop schemes in prototype set up. Experimental results show that this algorithm is simple and effective to control the tension even in transient state.

REFERENCES

[1] Chen Dechuan,Chen Zhilin.Winding Tension and Velocity Coordinated System with Torque Servo Control Mode[J].Journal of Textile Research,2009,30(6):118-121.

[2] ZHANG Hongchang，XI Anmin，LIU Hongfei，et al. Tension thickness adjustment efficiency in aluminum foil rolling ［J］. China Mechanical Engineering，2008，19(14): 1748 — 1750.

[3] WANG Xing， JIANG Yong . Research on copper fo-il post-processor tension control system. ［J］. Applied Mechanics and Materials，2011，43: 356 — 361.

[4] Vedrines M ,Gassmann V,Dominique K. Moving Web-tension Determination by Out-of-plane Vibration Measurements Using a Laser[J].IEEE Tr-ansaction on Instrumentation and Measurement,2009,58(1):207-213

[5] PENG Zhihui,Ma Guang,Zhong Hongming,etal.Study on Design and Compensation Algorithm for High Ac-curacy Tension Controller[J]. China Mechanical Engineering, 2009,58(1):207-213

[6] SHEN Long,ZHOU Junwei.Tension Control for Conti-nuous Hot Dip Galvanizing Line Baotou Steel[J].Science and Technology of Baotou Steel, 2013,39（4）：61-63.

[7] Wang Youzheng.Measuring and Calculation Method of Inertia Moment ofDecoiler[J].Electric Drive, 2013,43（10）：44-46.

[8] Yang Zewei,An Lianxiang,Wang Huifeng,Gao Lipeng.A New Calculation Method of Dynamic Torque of Coiler[J].Metallurgical Industry Automation,2009,33(3):66-68.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Similarity Model Based On Trend For Time Series

ShuaiFei Chen
College of Computer and Information
Hohai University
Nanjing, China
chenshuaifei163@163.com

Xin Lv*, Lin Yu, YingChi Mao, LongBao Wang, HongXu Ma
College of Computer and Information
Hohai University
Nanjing, China
*Corresponding author: lvxin.gs@163.com

***Abstract*—This paper presents a time series similarity matching model based on trend meeting the people's intuitive sense of trends characterize similarity. At the same time, the concept of similarity value is introduced in order to display the similarity of time series in a more intuitive form. In this model, the original time series are segmented according to the time series segmentation algorithm based on significant points. Each sub-section of the time series are mapped to a two-dimensional vector according to the slope and time span, and then symbolic the two-dimensional vector and calculate the distance between two time series of strings. Finally according to similarity calculation formula proposed, obtain the similarity value between the two time series. Experimental results show that the time series similarity matching model is good. In the aspect of similarity matching, the applicability, high efficiency.**

Key words-time series; data mining; similarity; trends; linear representation

I. INTRODUCTION

Similarity of time series [1] still does not have a uniform definition. It is a matter with a strong subjective on measuring the similarity of time series. A model is needed for judging similarity of time series. The factors that affect the similarity of time series are various, not just rely on the user's needs, but also on the purpose of the current mandate, etc. For example, currently similarity measure of time series are mainly based on the Euclidean distance [2] [3], based on dynamic time warping distance [4], as well as the largest common subsequence [5], etc. But these distance measurement methods have many insurmountable defects.

The advantages and disadvantages of various distance measures: (1)Euclidean distance is famous for calculating simple and used widely, can be applied to clustering [6] and classification [7] and other research areas. But its biggest shortcoming is that can only deal with the time series with the same length and sensitive to amplitude variation [8] of time series; (2) Dynamic time warping distance can overcome the shortcomings of the Euclidean distance, supporting dynamic bending time series. But this method is high time complexity [9], limiting its scope of application; (3) the longest common sub-sequence is too strict and sensitive to a certain local change, will lead to the similarity of being wrong estimate. Based on the shortage of various distance measurement, this paper chooses edit distance as the similarity measure criterion for time series.

This paper proposes a time series similarity matching model based on trend [10], dividing into three steps. First step, the original time series are segmented according to the time series segmentation algorithm based on important points, and then judging similarity of time series based on trends of each sub-time series, which can avoid over-reliance on each data point. Second step, combine the slope and time span of each sub-segment to process the fitting time series and map it to a two dimensional vector; Third step, symbol the time series in accordance with the value of the two components of the two-dimensional vector, and then calculate the corresponding strings edit distance, and finally the similarity of time series is calculated on the basis of pre-defined similarity formulas and output the result. According to a large number of experiments, it shows that the time series similarity matching model based on trend proposed calculation is simple and can be implement easily, is not sensitive to the time axis offset and noise date. Calculating the similarity by the edit distance also can avoid the potential problems which will be brought by the largest common subsequence method, while the model is suitable for the similarity calculation of unequal time series.

The rest of the paper is organized as follows: The second section introduces the linear representation of time series; the third section proposes the symbolic representation of time series; the fourth section introduces the similarity matching algorithm of time series; the fifth section, results and discussion; section VI, conclusion.

II. TIME SERIES PIECEWISE LINEAR REPRESENTATION AND RELATED DEFINITIONS

This paper selects piecewise linear representation method based on significant points to preprocess the original time series. The original series are segmented according to import points. The purpose is to remove the details of interference, only preserve the main morphological features of the time series, which will be helpful to improve the efficiency and accuracy of data mining [11].

Definition 1 (important point): suppose there are time series $X =< x_1, x_2, \cdots, x_n >$, $X_{ij} =< x_i, x_{i+1}, \cdots, x_j >$ is sub-time series X. When the point a_m satisfy one of the following conditions, then a_m is an important point,

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.115

1) a_m is a minimum value, while $a_i / a_m \geq R, a_j / a_m \geq R$;

2) a_m is a maximum value, while $a_i / a_m \leq R, a_j / a_m \leq R$.

Definition 2 (time series piecewise linear representation) suppose there are time series $X =< x_1, x_2, \cdots, x_n >$, segmentation points set of X is $X'_i = <x'_1, x'_2, \cdots x'_m>$, where $x'_1 = x_1, x'_m = x_n, m < n$, the piecewise linear representation of X is:

$$X_L =< f_1(x'_1, x'_2), f_2(x'_2, x'3), \cdots; f_{m-1}(x'_{m-1}, x'_m) > \quad (1)$$

Where $f_{m-1}(x'_{m-1}, x'_m)$ represents the linear fit within the range of $[x'_{m-1}, x'_m]$ functions. Definition 3 (fitting errors) suppose the time series $X =< x_1, x_2, \cdots, x_n >$, segmentation points set X_L is obtained by segment the time series X., after linear interpolation, the X_L referred to as fitting sequence X^C, $X^c =< x_1{}^c, x_2{}^c, \cdots x_n{}^c >$. The fitting sequence and the original time series' fitting error are:

$$E = \sqrt{\sum_{i=1}^{n} (x_i - x_i{}^c)^2} \quad (2)$$

The time series is segmented and the segmented points set is obtained by using the piecewise linear representation method based on important points. Then using the linear interpolation method to connect the adjacent segmented point, so the piecewise linear representation of the original time series is gotten. In order to guarantee the validity of the experimental results, the fitting error between the fitting time series and the original time series is need to be controlled within a certain range.

III. SYMBOLIC REPRESENTATION OF TIME SERIES

Linear fitting of time series by extracting important point, the main features of the original time sequence is preserved, showing a series of rising and falling trend. In order to quantify the similarity between time series, discrete and symbolic processing is necessary for fitting time series. Depending on the time sequence of rise and decline, you can use different degrees of finely divided (rapid rise, rise slowly, declines rapidly, slow decline, etc.), then symbolize.

TABLE I. SYMBOLIC RULES

Changing Trend	**Time Span**	
	Greater than Q	***Less than or equal to Q***
Dramatic rise	A	B
Slow rise	C	D
Sharp decline	a	b
Slow decline	c	d

However, if only in accordance with the rise or decline trend of the time series to classify sub-segments of the time series has a problem: the time spans for each sub-segment are different, which may lead to erroneous estimation of similarity, as shown in Figure 1. Therefore, symbolic the time series, the time span needs to be taken into account.

In this paper, we take the following symbolic strategy: combine trends and time span of the time series' sub-segments are classified, each category corresponds to a symbol of the rule, and the specific classification rules could be controlled according to user needs. Assumed trends divided into four levels, the time span is divided into two levels, the corresponding symbol rules is shown in Table 1.

Cai Zhi, who proposed a similarity measure method based on the slope of the tangent value. But this method ignores the time span's effect on time series similarity, so the situation in Figure 4-2 will be judged error. The method which combines the trend and the time span proposed in this paper can avoid the misjudgment.

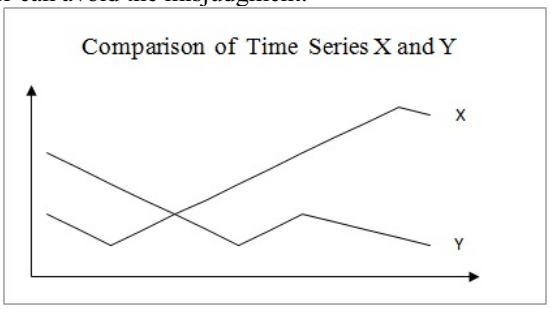


Figure 1. Ignore time span

IV. TIME SERIES SIMILARITY MATCHING ALGORITHM

A. *relevant definitions*

Definition 4 (edit distance) edit distance means between two strings, made a turn into another minimum number of operations required for editing. License editing operations include the substitution of one character to another character; insert a character; deleting a character.

After the symbolic processing of the original time series, we obtain the trend strings S_X and S_Y. Calculate the edit distance between S_X and S_Y. The difference of the two strings can be obtained.

Definition 5 (time series similarity degrees) assume that the time series X and Y, we can get fit sequence X_L and Y_L through processing the original time series according to the method of piecewise linear representation. Then symbol X_L and Y_L, get trends string S_X and S_Y. The similarity of two time series is calculated as follows:

$$sim(\mathrm{X}, \mathrm{Y}) = \left(1 - D_{edit} / \max(L(S_X), L(S_Y))\right) * \left(\frac{\min(L(X), L(Y))}{\max(L(X), L(Y))}\right) * \left(\frac{\min(\Delta_{Amp}(X), \Delta_{Amp}(Y))}{\max(\Delta_{Amp}(X), \Delta_{Amp}(Y))}\right) \quad (3)$$

$L(X)$, $L(Y)$ represents length of the original time series and $\Delta_{Amp}(X)$, $\Delta_{Amp}(X)$ is the amplitude span of the original time series. $L(S_X)$, $L(S_Y)$ is the length of the corresponding

symbol string. According to different application scenarios, the amplitude of the formula span is optional.

B. Similarity calculation method based on time series trend

Algorithm steps are as follows:
Input: Time Series X and Y
Output: Similarity between X and Y

a) Step One: According to the time series segmentation algorithm based on significant points linear piecewise representation method, obtain sub-point set X' and Y'. Then connecting the two adjacent important point by means of linear interpolation, so we get the original time series piecewise linear representation X_L and Y_L;

b) Step Two: firstly, linear fit every time series' sub-segment mapped to a two-dimensional vector $X_{subi} = (K_{subi}, L_{subi})$, where K_{subi} and L_{subi} represent the slope and the time span of the first i sub-segment. The slope of the time series of the time series is divided into M sub-segment species categories, and the time span of the sub-segment is divided into N sub-segments categories, so that time series' sub-segment is divided into $M \times N$ sub-segment species categories corresponding to $M \times N$ types of symbol rules. Two time series respectively correspond to the rules of symbolic, resulting in a corresponding trend series S_X and S_Y;

c) The third step: Calculate trends string edit distance between S_X and S_Y, obtaining $sim(X,Y)$. Then calculate the similarity of two time series and output according to Equation 3.

V. Results and Discussion

In order to verify the validity of the model proposed in this paper, selected time series in different fields for verification. Experiment was divided into two parts: The first part, given a time sequence specified pattern information of interest, quickly and accurately retrieve all the time series of sub-segments in line with the trend given by the model mode; Part II: Given two approximately equal length of time series, calculate the degree of similarity between the two time series, and gives a quantitative representation.

A. Experiment one

The purpose of the experiment is to retrieve all sub-time series having the same pattern. The experimental time series dataset are from www.cs.ucr.edu/~eamonn/tutorials.html, universal Time series dataset for time series data mining (herein referred to as KData), as shown in Table Ⅱ.

First, for a given time series, Symbolic it through symbolic strategy of the model, and then use the string search techniques to find out the number of sub-segments that meet the specified time series mode. Table Ⅲ is the number of sub-segments statistics which satisfies doublet mode (specified for each rising or falling trend time span of at least 10 basic time unit).

TABLE II. KDATA

Sequence Name	Length	Sequence Name	Length
Burst	9382	Memory	6875
Chaotic	1800	Ocean	4096
Fluid_dynamics	10000	Earthquake	2501
Leleccum	4320	Tide	6985

TABLE III. NUMBER OF BIMODAL PATTERN SUB-SEGMENT STATISTICS

Sequence Name	Total Number	Sequence Name	Total Number
Burst	38	Memory	8
Chaotic	22	Ocean	2
Fluid_dynamics	10	Earthquake	1
Leleccum	7	Tide	4

According to the results, after symbolic time series, we can use a string search technology search out all the similar sub-time series quickly and accurately. Because the string representation is the morphological characteristics of time series, so the model is not sensitive for noise date, offset on the time axis and amplitude changes. When dividing the time span and trends finer, the similarity result of time series is more accurate.

B. Experiment two: calculate two given time

The experiment is to computing the similarity value of two given time series by similarity model proposed in the paper, verifying the validity of the model. From the results of the first experiment, select first two sub-time series of Burst and Chaotic that meets the bimodal pattern as shown in Figure 2. Calculate the similarity and output specific similarity value.

TABLE IV. SIMILARITY RESULT STATISTICS

Candidate time series	Similarity Value	
	Consider amplitude span	***Irrespective of amplitude span***
（Burst:1,Burst:2）	91.5%	75.3%
(Chaotic:1,Chaotic:2)	99.53%	98.2%

According to the time series graphs Figure 2 and experimental results Table Ⅳ, we can get specific time-series similarity value between time series. At the same time, we can see that the time series similarity matching model based on the trend are not sensitive for noises data, offset on the time axis. So the model is suitable for varying time series similarity calculations.

Depending on the specific application scenarios, When calculating the similarity of time series, the amplitude span as an option (for example, voice recognition is not considered amplitude). When judging that whether two time series are similar, the similarity threshold can be set in advance. If the similarity value is greater than the threshold value, it is considered that the two time series are similar.

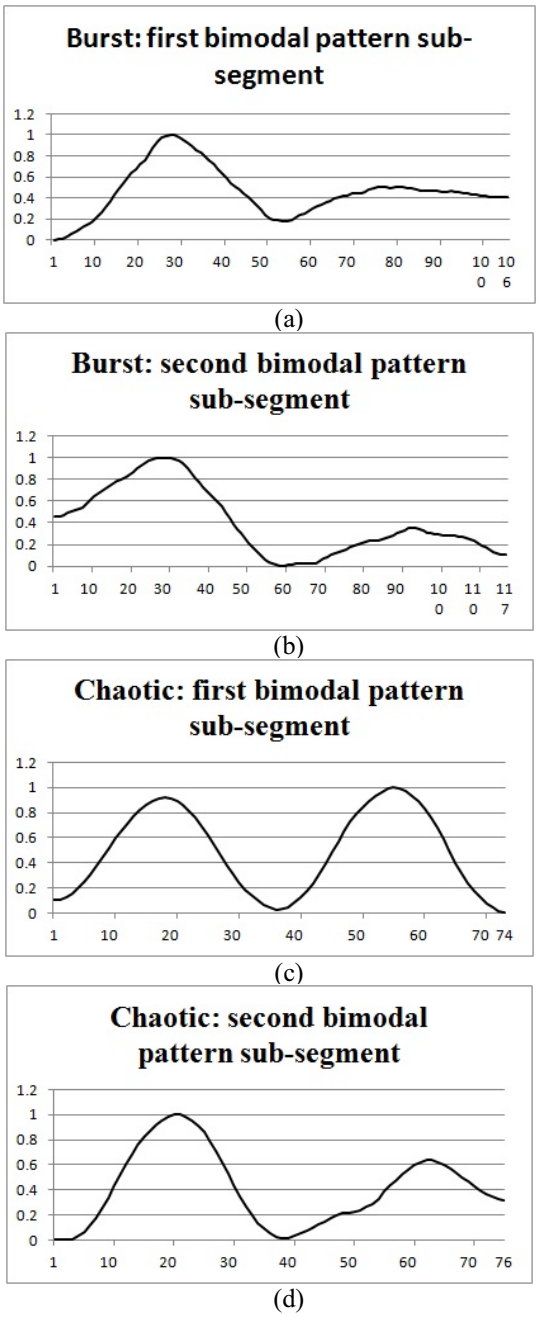


Figure 2. THE FIRST TWO BIMODAL PATTERN SUB-SEGMENT OF BURST AND CHAOTIC

VI. CONCLUSION

The similarity model for time series based on trend proposed in this paper integrated representation of time series, symbolic and similarity measure in a single framework. Conceptually, the model is in line with people's way of thinking in the observation of graphics to construct by the major trends and important point of the graph. The model can calculate the similarity value accurately of time series, and different sizes can be taken to extract trends according to the different needs of users.

Because of the time series similarity matching model proposed is based on morphological characteristics of time series to compare similarities, so the model for noises, offset on the time axis are not sensitive, while the model is also suitable for long range time series the similarity comparison. The next phase of the research is to consider how to make the trend of the time series more reasonable, and to improve the accuracy of the similarity matching of time series based on the trend.

ACKNOWLEDGMENT

This research is partially supported by "National Natural Science Foundation of China" (Grant No. 61272543); "National Key Technology Research and Development Program of the Ministry of Science and Technology of China" (Grant No. 2013BAB06B04); "Key Technology Project of China Huaneng Group" (Grant No. HNKJ13-H17-04); "Natural Science Foundation of Jiangsu Province" (Grant No. BK20130852); "Jiangsu Planned Projects for Postdoctoral Research Funds" (Grant No. 1401001C).

REFERENCES

[1] Keogh E, Chakrabarti K, Pazzani M, et al. "Dimensionality reduction for fast similarity search in large time series databases" [J]. Knowledge and information Systems, 2001, 3(3): pp.263-286.

[2] Wu Y L, Agrawal D, El Abbadi A. "A comparison of DFT and DWT based similarity search in time-series databases," [C] Proceedings of the ninth international conference on Information and knowledge management. ACM, 2000: pp.488-495.

[3] Chan K P, Fu A W C. "Efficient time series matching by wavelets," [C] Proceedings of the 1999 15th International Conference on Data Engineering, ICDE-99. IEEE, 1999: pp.126-133.

[4] Bernad D J. "Finding patterns in time series: a dynamic programming approach,"[J] Advances in knowledge discovery and data mining, 1996.

[5] Bollobás B, Das G, Gunopulos D, et al. "Time-series similarity problems and well-separated geometric sets," [C] Proceedings of the thirteenth annual symposium on Computational geometry. ACM, 1997: pp.454-456.

[6] Liao T W. "Clustering of time series data—a survey," [J]. Pattern recognition, 2005, 38(11): pp.1857-1874.

[7] Ongenae F, Van Looy S, Verstraeten D, et al. "Time series classification for the prediction of dialysis in critically ill patients using echo statenetworks," [J] Engineering Applications of Artificial Intelligence. Elsevier Ltd, 2013, 26(3): pp.984-996.

[8] Perng C S, Wang H, Zhang S R, et al. "Landmarks: a new model for similarity-based pattern querying in time series databases," [C] 2000 IEEE 16th International Conference on Data Engineering (ICDE'00). IEEE, 2000: pp.33-42.

[9] Yi B K, Jagadish H V, Faloutsos C. "Efficient retrieval of similar time sequences under time warping," [C] Data Engineering, 1998. Proceedings. 14th International Conference on. IEEE, 1998: pp.201-208.

[10] Pratt K B, Fink E. "Search for patterns in compressed time series," [J] International Journal of Image and Graphics, 2002, 2(01): pp.89-106.

[11] Z. Zhou, M Q Li. "Time series segmentation based on series important points," [J] Computer Engineering, 2008, 34 (23): pp.14-16.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A Hill-type Submaximally-activated Musculotendon Model and Its Simulation

Lixin Sun, Yingfei Sun, Zhipei Huang, Jiateng Hou, Jiankang Wu
Sensor Network and Application Research Center(SNARC)
University of Chinese Academy of Sciences(UCAS), Beijing, China
E-mail: solotraveller89@gmail.com yfsun@ucas.ac.cn

***Abstract* - Hill-type models are ubiquitous in biomechanical simulations because of their computational simplicity and efficiency. But these models are designed to describe the maximally activated muscles and muscle properties are linearly scaled when they are applied to submaximally activated conditions. This scaling approach should be based on the independence of muscle activation and force-length properties, which has not been proven yet. Actually, muscles in vivo are unlikely to be often maximally activated during daily life. Therefore, effective methods should be taken to modify the existing Hill-type model to insure the accuracy of their applications. This paper analyzed the submaximal activation conditions, developed a submaximally-activated musculotendon model on the basis of Millard damped equilibrium model and implemented the benchmark experiments to verify effectiveness of the modified model.**

Keywords - Hill-type model; musculotendon model; muscle contraction dynamics; submaximal activation

I. INTRODUCTION

Estimation of individual muscle forces during human movement can provide insight into neural control and tissue loading and can thus improve diagnosis and management in both neurological and orthopaedic conditions. Unfortunately, direct measurement of muscle forces in vivo is not feasible in a clinical setting[1]. Therefore a mathematical musculotendon model, which takes anatomical and physiological characteristics of muscles into account is required. There are two broad classes of musculotendon models: Huxley-type molecular (i.e., cross-bridge) models and Hill-type phenomenological models[2]. Hill-type models are widely used in muscle-driven simulations of human and animal motions because of their computational simplicity and close relation to measured experimental variables[3].

The original Hill-type model is developed by English physiologist A.V. Hill[4], who measured the heat produced by muscle during contraction and concluded the force-velocity relation equation, i.e., the famous Hill equation. Meanwhile, Hill proposed a two-component model for the mechanical behavior of muscle, consisting of an undamped purely elastic element called “series elastic element(SEE)”, in series with a “contractile element(CE)”, which is governed by the characteristic Hill equation. Later, a third “parallel elastic element(PE)” is added to the model considering the fact that passive, unactivated muscle can resist stretches. Thus PE with the original two elements, SEE and CE, constitutes the classical three-element Hill muscle model. A more detailed introduction to the Hill model is described by Zahalak[4] and Winter[5].

However, Hill didn't take tendon elasticity effects and muscle activations into account. Zajac[6] completed the contraction dynamics model and formed a one-parameter model, which takes muscle and tendon as one entity(the musculotendon actuator). Zajac's model neglected SEE, which is one element of the Hill model, because SEE stores little energy compared to the tendon. And this practice conformed to the basic notion that sarcomeres and fibers act in concert. Zajac offered not only a functional model of musculotendon actuator, but also a general mathematical model. But this model doesn't take pennation angle into account and will cause a singularity problem when force-length-velocity relation is not invertible, activation is low or maximum isometric force is small. Zajac has resolved this problem by constraining the parameters with lower bounds.

Based on Zajac's work, Schutte[7] built a muscle contraction dynamics model, which takes the muscle fiber length as the state variable instead of tendon force to avoid the singularity when inversing tendon force-length curve. To solve the original singularity problem, Schutte included a nonzero parallel passive damping element in the muscle part and avoided the singularity perfectly.

Also based on Zajac's work, Lloyd et al.[8] constructed a model adding the effect of pennation angle and acquired length-varying pennation angle by assuming that muscle has constant thickness and volume as it contracts. In addition, Lloyd also took the coupling between muscle activation and fiber length into consideration and proved an improvement in model accuracy. Moreover, an EMG-driven musculoskeletal model to estimate muscle forces and knee joint moments in vivo was built based on this muscle contraction dynamics model and worked well.

Thelen et al.[9] combined the model of Zajac's and Schutte's and complemented them in the aspect of muscle property curves, like active force-length relationship, passive force-length relationship and so on by function fitting them with appropriate parameters. Parameters sensitive to ages were also pointed out by experiments. In the end, Thelen et al. integrated this model into human motion equations and acquired muscle activation, fiber length and kinematics parameters at each moment by static optimization algorithms, i.e. the CMC(Computed Muscle Control) tool in OpenSim[10][11], an open-source software system for musculoskeletal dynamics analyzation. Actually, Chalfoun et al.[12], Csercsik[13] and Lemos et al.[14] used different functions to fit muscle property curves, but influences on the accuracy of the model are not explored yet.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.116

Subsequently, Millard et al.[2] implemented three musculotendon models: an equilibrium model, a damped equilibrium model and a rigid tendon model. And benchmark simulations were designed to compare their speeds and accuracies. All the three models have been implemented in OpenSim and share the same muscle property curves, i.e., quintic Bezier splines fitting to experimental data.

Generally, Hill-type models are designed to describe the maximally activated muscle only and muscle properties are linearly scaled when they are applied to submaximally activated conditions. Experimental validations of these models by Perreault et al.[3] revealed that model errors during movements are largest at low motor unit firing rates relevant to normal movement conditions. As during daily activities, muscle is not likely to be maximally activated very often, a model appropriate for submaximal activations needs establishing to assure the accuracy of muscle-driven simulations of human normal movements.

The purpose of this paper is to introduce a submaximally-activated model, which is built on the basis of the damped equilibrium model developed by Millard et al.. In Section II, Millard damped equilibrium model will be detailed and a modification method will be introduced after submaximal activation conditions being analyzed. Then experiments will be described and experimental results will be analyzed and discussed in Section III.

II. SUBMAXIMALLY-ACTIVATED MUSCULOTENDON MODEL

In order to maintain the integrity of the proposed model, the original Millard damped equilibrium model will be described in more detail. Then submaximal activation conditions are analyzed. Finally, a feasible modified method will be proposed.

A. Millard damped equilibrium model

As shown in Figure 1, the model consists of an active contractile element, a passive elastic element, a parallel damping element and an elastic tendon. As the mass of the muscle is assumed to be negligible, the muscle and tendon force should be in equilibrium, then the musculotendon model can be described by (1).

$$F^{max}[af_L(\tilde{l}_m)f_V(\tilde{v}_m)+f_P(\tilde{l}_m)+\beta\tilde{v}_m]cos(\alpha)-F^{max}f_T(\tilde{l}_t)=0 \quad (1)$$

where maximum active force-length curve $f_L(\tilde{l}_m)$, passive force-length curve $f_P(\tilde{l}_m)$, tendon force-length curve $f_T(\tilde{l}_t)$ all represent forces normalized by maximum isometric force F^{max}; $\tilde{l}_m$, $\tilde{v}_m$ and $\tilde{l}_t$ represent muscle length, muscle velocity and tendon length normalized by optimal fiber length, maximum velocity and tendon slack length, respectively. β is damping coefficient, α is pennation angle, a is muscle activation ranging from $a_{\min}$ to 1.

The Millard damped equilibrium model is a comparatively mature version of Hill-type model because it has solved the singularity problem with a damp element and boundary constraints, which avoid onerous calculations during the process of numerical integration. The four

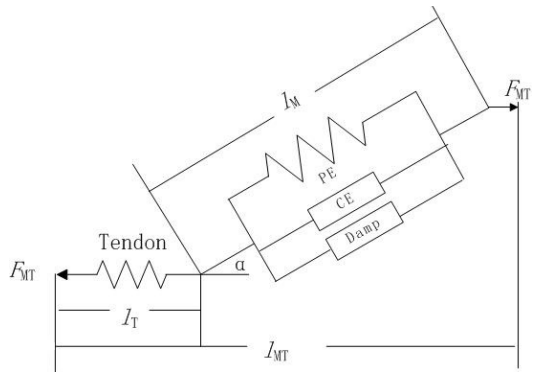


Figure 1 Millard damped equilibrium musculotendon model

musculotendon property curves are developed using Bezier splines with default control points fitting to experimental data. Bezier splines are C_2-continuous (i.e., continuous to the second derivative) and straightforward to modify, so it's easy to adjust the curves to different kinds of muscles according to their characteristics. Besides, this model is implemented in OpenSim, an open-source and widely used in biomechanics simulations.

Nonetheless, as an inherent limitation of Hill-type model, Millard damped equilibrium model also has larger errors at submaximal activations than maximal ones. Probable causes of this phenomenon will be analyzed and demonstrated below.

B. Submaximal activations analysis

Huijing[15] indicated that the most limiting factor of the phenomenological models is that they are designed to describe effects in maximally activated muscles only. And the reason why submaximally activated situations are not paid enough attentions is that most experiments describing muscle properties are performed under conditions of supramaximal stimulation of the peripheral nerve. But muscles are not likely to be maximally active very often during daily activities, so this problem must be dealt with if the model is to be applied to the musculoskeletal system of human movement.

The general way to deal with submaximally activated situations is to scale down the force-length curve with normalized activations. This linear scaling approach is under two hypotheses: (a) The relationship between the force and the normalized activation is linear. (b) The length is independent of activation. As for a fully recruited muscle, submaximally activation means manipulating the firing rate, then there should be a linear force-firing rate relationship. Skinned muscles can be directly activated by the external application of Ca^{2+} buffered solutions and changes in intracellular Ca^{2+} concentration can be monitored by suitable Ca^{2+} indicators. Therefore the intracellular equivalent relationship of force-firing rate can be achieved by experiments on skinned muscle fibers as the force-pCa ($-\log_{10}[Ca^{2+}]$) curve. Stephenson and Wendt[16] presented a few isometric force-pCa curves obtained with different types of mechanically skinned skeletal muscle fibers. These curves are clearly sigmoidal in shape, which breaks the first hypothesis. And the shape of force-pCa curve vary with

muscle length, as the apparent sensitivity of the contractile proteins to Ca^{2+} depend on muscle length, then the second hypothesis fails.

When coming back to the force-length curve, the direct observation is shift in the plateau toward longer lengths as the level of activation decreases. Experiments on both sarcomeres and intact fibers of different types of muscles have shown the similar results. In addition, Huijing[15] indicated the feature that both active slack length and optimum length shift substantially to higher muscle length at lower constant frequencies and the shift distance of the former is smaller than the latter. And the maximal length of force exertion is considered a fixed property of muscle regardless of its degree of activation.

C. Modified method

Actually, Huijing only qualitatively described the coupling between muscle activation and fiber length described above and appealed to muscle modellers to take this property into account. Hatze[17] incorporated effects of firing frequency in his model by altering the slope of the force-frequency curve as a function of muscle length, but didn't fully succeed for ignoring shifts of active slack length. Van Zandwijk et al.[18] modified Hatze's model and excelled in predicting muscle forces by optimizing the model for an isometric twitch. However, both the models take Ca^{2+} concentration as one input, which is only measured in the special experiments meaning that the models are not appropriate to integrate to the musculoskeletal model for human movement analysis.

When constructing an EMG-driven musculoskeletal model of the human knee, Lloyd[8] built a muscle contraction dynamics model, which incorporated the coupling between activation and optimal fiber length as (2).

$$L_m^o(t) = L_m^o(\gamma(1-a(t))+1) \qquad (2)$$

where γ is the percentage change in optimal fiber length, L_m^o is optimal fiber length at maximum activation, $a(t)$ and $L_m^o(t)$ are activation and optimal fiber length at time t, respectively. This equation changes the force-length curve indirectly and shifts the whole curve to larger lengths. Lloyd examined the effect of the physiological parameter γ in a contrast test with the knee model, which predicted joint moments in vivo and revealed that the joint moment R^2 significantly increase 0.6 by including γ whose default value is 0.15.

In consideration of Lloyd's successful modification and advantages of Millard damped equilibrium model, a possible improvement method can be designed. From aspect of the optimal fiber length, just the same as Lloyd's, integrate (2) into the Millard model and setting aside an adjustable γ factor.

III. Biological benchmark experiment, results and Discussion

A. Biological benchmark experiment

Millard et al.[2] offered a set of submaximal-activation biological benchmark experiment, which contains experimental data on cat soleus muscle in vivo. There are three groups of experimental trials and the first group includes six trials, in which the musculotendon actuator is held a constant length and excited using constant-frequency stimulation rates of 10, 20 and 30 Hz and random signals with mean frequencies of 10, 20 and 30 Hz. The other two groups both have six trials and share the same mode of stimulations with the first one, but apply length changes with maximum amplitudes of 1.0 mm and 8.0 mm respectively to the free end of the tendon. The data of the first isometric group is used to calculate the activation signals of corresponding stimulations and these activations are then applied to the following two non-isometric groups. Experimental details and muscle parameters can be found in Perreault et al.[3] and Millard et al.[2].

As the modified model incorporates the coupling relationship between muscle activation and length, it's complicated to calculate activations from the measured forces, so activations calculated from the original model are adopted. Then the same activations and time-varying musculotendon lengths are input to the modified model. Errors between the experimentally measured muscle forces and those predicted by the two models are quantified using percent root mean square (RMS) values.

B. Results

The RMS errors of the modified model using different optimal γ values in the two groups of experimental trials are shown in Figure 2, compared with those of the original one. Generally, errors of the 1.0 mm displacement trials are much smaller than the 8.0 mm ones. The modified model performed better than the original one and exceled the original one in all the trials except the Random 10 Hz one in 1.0 mm displacement group. The accuracy of the modified model is γ-dependent and the 1.0 mm displacement group prefers smaller values. To summarize, errors of the original model are less than 3.35% and 16.59% corresponding to displacements of 1.0 mm (top row in Figure 2) and 8.0 mm (bottom row), respectively. Meanwhile, those of the modified model are less than 2.58% (when γ=0.25) and 12.25% (when γ=0.45), correspondingly.

C. Discussion

In fact, the activations, which are input of the simulations, are calculated using the original model, results should be better if they can be calculated using the modified one. The particular case when the modified model performs worse may be caused by this reason. This modification method is easy to implement and moves the whole curve to the right indirectly, but ignores the change of distance between the active slack length and the optimum fiber length and property of fixed maximum fiber length. A direct modification to the property curve is implemented, but more time is cost without obvious improvement. So this modification is recommended to be used in the construction of upper-layer musculoskeletal system.

This paper verified the dependence of length and activation from the perspective of applications, tested a simple and effective way to incorporate the dependence into

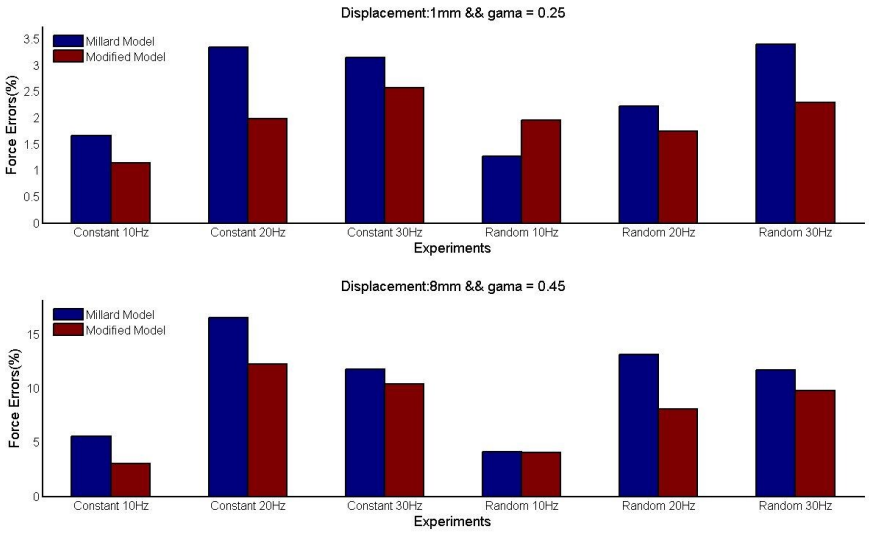


Figure 2 Comparison of the root mean square (RMS) errors of two musculotendon models for submaximally-activated muscle undergoing length changes of maximum amplitude 1.0 mm(top) and 8.0 mm(bottom)

Hill-type model and constructed a practicable submaximally-activated musculotendon model. Yet the RMS errors still need improving, future work can take factors like contracting history into consideration to improve the accuracy. More improvement methods should be devoted to the study of musculotendon models, especially the widely-used Hill-type model.

ACKNOWLEDGMENT

The authors wish to acknowledge the support of the National Natural Science Foundation of China (61431017) and the assistance of Thomas Uchida in implementing the benchmark experiment.

REFERENCES

[1] A. Erdemir, S. McLean, W. Herzog, and A. J. van den Bogert, "Model-based estimation of muscle forces exerted during movements," Clin. Biomech., vol. 22, no. 2, pp. 131–154, 2007.

[2] M. Millard, T. Uchida, A. Seth, and S. L. Delp, "Flexing computational muscle: modeling and simulation of musculotendon dynamics.," J. Biomech. Eng., vol. 135, no. 2, p. 021005, 2013.

[3] E. J. Perreault, C. J. Heckman, and T. G. Sandercock, "Hill muscle model errors during movement are greatest within the physiologically relevant range of motor unit firing rates," J. Biomech., vol. 36, no. 2, pp. 211–218, 2003.

[4] G. Zahalak, "Modeling Muscle Mechanics (and Energetics)," in Multiple Muscle Systems SE - 1, J. Winters and S.-Y. Woo, Eds. Springer New York, 1990, pp. 1–23.

[5] J. Winters, "Hill-Based Muscle Models: A Systems Engineering Perspective," in Multiple Muscle Systems SE - 5, J. Winters and S.-Y. Woo, Eds. Springer New York, 1990, pp. 69–93.

[6] F. E. Zajac, "Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control.," Crit. Rev. Biomed. Eng., vol. 17, no. 4, pp. 359–411, 1988.

[7] L. M. Schutte, "Using musculoskeletal models to explore strategies for improving performance in electrical stimulation-induced leg cycle ergometry." Stanford University, 1992.

[8] D. G. Lloyd and T. F. Besier, "An EMG-driven musculoskeletal model to estimate muscle forces and knee joint moments in vivo," J. Biomech., vol. 36, no. 6, pp. 765–776, 2003.

[9] D. G. Thelen, "Adjustment of muscle mechanics model parameters to simulate dynamic contractions in older adults.," J. Biomech. Eng., vol. 125, no. 1, pp. 70–77, 2003.

[10] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "OpenSim: open-source software to create and analyze dynamic simulations of movement," Biomed. Eng. IEEE Trans., vol. 54, no. 11, pp. 1940–1950, 2007.

[11] A. Seth, M. Sherman, J. a. Reinbolt, and S. L. Delp, "OpenSim: A musculoskeletal modeling and simulation framework for in silico investigations and exchange," Procedia IUTAM, vol. 2, pp. 212–232, 2011.

[12] J. Chalfoun, R. Younes, M. Renault, and F. B. Ouezdou, "Physiological Muscle Forces, Activation and Displacement Prediction During Free Movement in the Hand and Forearm," J. Robot. Syst., vol. 22, no. 11, pp. 653–660, 2005.

[13] D. Csercsik, "Analysis and Control of a Simple Nonlinear Limb Model," 2005.

[14] R. R. Lemos, M. Epstein, W. Herzog, and B. Wyvill, "A framework for structured modeling of skeletal muscle.," Comput. Methods Biomech. Biomed. Engin., vol. 7, no. 6, pp. 305–317, 2004.

[15] P. a. Huijing, "Muscle, the motor of movement: Properties in function, experiment and modelling," J. Electromyogr. Kinesiol., vol. 8, no. 2, pp. 61–77, 1998.

[16] D. G. Stephenson and I. R. Wendt, "Length dependence of changes in sarcoplasmic calcium concentration and myofibrillar calcium sensitivity in striated muscle fibres," Journal of Muscle Research and Cell Motility, vol. 5, no. 3. pp. 243–272, 1984.

[17] H. Hatze, "A myocybernetic control model of skeletal muscle," Biol. Cybern., vol. 25, no. 2, pp. 103–119, 1977.

[18] J. P. van Zandwijk, M. F. Bobbert, G. C. Baan, and P. A. Huijing, "From twitch to tetanus: performance of excitation dynamics optimized for a twitch in predicting tetanic muscle forces," Biol. Cybern., vol. 75, no. 5, pp. 409–417, 1996.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Evaluation of Testing Software Program Based on DEA with Fuzzy Window

LI Xiang-Juan
College of Business Administration,
Nanjing University of Traditional Chinese Medicine
Nanjing, PRC
1784449598@qq.com

KE Zun-You
Information Engineering Dept.,
Nanjing Institute of Mechatronic Technology,
Nanjing, PRC
clickyouyou@126.com

***Abstract*—Based on the DEA (Data Envelopment Analysis) traditional model, an improved evaluation model is introduced for the fuzzy index value and fuzzy preference weight information. The improved model is applied in evaluation of automatic testing software programs. Firstly, the subjective index values are transformed into fuzzy numbers. Secondly, the subjective preference weights are constructed as a constrain of fuzzy window. Finally, the improved fuzzy DEA model is established to access the comprehensive evaluation of automatic testing software programs. A typical solving algorithm for the evaluation model is practiced. The instance simulation carries out the evaluation with direct subjective weight preference information, resulting in automatic testing software programs keeping consistence for different confidence levels.**

Keywords- system modeling; data envelopment analysis; fuzzy window; testing software program evaluation

I. INTRODUCTION

Software testing is an important part of the software development process. Software enterprises try hard to reduce the cost of software testing and enhance testing efficiency. At present, it's quite common that the automatic testing software tools are employed to improve the efficiency, effectiveness and quality.

From 80s in the last century, a number of studies focus on automatic testing software to improve test results, but in the face of increasingly complex and diverse testing software choices. Even to expert assessments, it is also difficult to determine the precise indicators for each index value of automatic testing software. Subjective language or fuzzy quantitative forms are frequently used in discussions. And due to the complexity and vagueness of description, it's also difficult to give the exact weight of each index. Usually some subjective quantitative or qualitative relationships between weights can be accessed [1]. Therefore, it's quite valuable to research questions about the comprehensive evaluation of automatic testing software tools based on subjective values.

Since these index values are subjective, so fuzzy theory evaluation process is applicable, such as evaluation methods of fuzzy comprehensive evaluation and fuzzy analytic hierarchy process [2][3]. Researches and evaluation methods are given by some scholars for index values as fuzzy numbers in evaluation, in which generally the specific weight of each index should be determined in advance [4][[5]6], and the weight information is subjective. Some papers give the corresponding evaluation methods to both indexes and weights are fuzzy [7][8], however all these methods do not process the following:

- The data structure is based on a specific scenario without various types of fuzzy numbers;
- The weights of each index need to be prepared without consideration of incomplete information on weights;
- During uncertain weight information and fuzzy DEA (Data Envelopment Analysis) evaluations, the certain weight linear information [9][10], but no weights of fuzzy preference relations is concerned.

DEA improved evaluation methods introduced here do not need weight information in advance. The improved fuzzy DEA model is introduced to resolve that indicator values are fuzzy and weights are fuzzy. The software testing management and decision-making capabilities are enhanced.

II. DATA AND METHODOLOGY

A. Evaluation Index

International organization for standardization (ISO) published a series of software product quality standards, which contains features defining software quality characteristics and properties[11], as Table I.

TABLE I. SOFTWARE QUALITY FEATURES

Features	Sub-features
Functionality	Adaptability, accuracy, interoperability, security, functionality compliance
Reliability	Maturity, fault tolerance, recoverability, reliable compliance
Usability	Understandability, learnability, operability, attractiveness, and usability compliance
Efficiency	Time, resource utilization and efficiency compliance
Mantainability	Analyzability can change, stability, testability, and maintain compliance
Portability	Adaptability, installability, alternative, coexistence, compliance of transplantation

In practice, testing software evaluation indexes are mapped to these six features, and subsequent data based on these categories are applied in the fuzzy evaluation analysis. The index and weight parameters are expressed as fuzzy numbers.

B. Index Fuzzification

Suppose there are n optional automation testing software programs and each has m indexes. The overall qualitative values are recorded as $r_{ij}\ \ (i = 1,2,...,m;\ j = 1,2,...,n)$. These indexes are usually in a form of subjective evaluation.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.117

By using fuzzy theory, the linguistic assessments or subjective evaluations in the form of quantitative analysis are transformed to qualitative indicators fuzzy number in interval [0,1]. The five language values "very poor", "poor", "middle", "good", "excellent" can be converted to fuzzy numbers respectively as: "very poor": (0,0.3); " poor ": (0.2, 0.5);"middle": (0.4,0.7);" good ": (0.6,0.9);" excellent ": (0.8,1.0).

C. Preference Weight Fuzzification

There are two types of subjective preference weights fuzzification , by the direct or comparative weight.

(A) The direct subjective weight can be transformed into interval number. In quantitative analysis, linguistic qualitative indicators values are converted to the fuzzy interval [0,1]. E.g. "not concerned" or "tightly concerned", "concerned", "concerned more", "concerned most", five language values can be converted into the interval fuzzy number: "not concerned": (0,0.3); "tightly concerned": (0.2,0.5); "concerned": (0.4,0.7); "concerned more": (0.6,0.9); "concerned most": (0.8,1.0).

(B) The comparative fuzzy preference can be transformed into fuzzy sequence information. In quantitative analysis, the comparative qualitative indicators are converted into the fuzzy interval [0,1]. E.g. "more or less", "a little more important", "more important", "much more important" and "most important privilege " five language value of the language variable are accordingly converted to interval fuzzy number (0,0.3); (0.2,0.5); (0.4,0.7); (0.6,0.9) and (0.8,1.0).

D. Cut Set for Solution

As fuzzy interval $r_{ij} = (r_{ij1}, r_{ij2})$, here $r_{ij1} \leqslant r_{ij2}$, r_{ij} represents the fuzzy qualitative value of the optional program j $(j = 1,2,...,n)$ on index i, and r_{ij} is among real interval [0,1]. The bigger r_{ij} is, the better attribute of program j is on index i.

The cut set of the fuzzy number is used to solve the fuzzy planning problems.

Supposed confidence level $\alpha(0 \le \alpha \le 1)$, i.e. $(r_{ij})_\alpha = \min\{r_{ij} \mid r_{ij} \in$ supp (r_{ij}) ,and $\mu(r_{ij}) \geqslant \alpha\ \}$, then $(r_{ij})_\alpha = r_{ij1} + \alpha(r_{ij2} - r_{ij1})$. Especially $(r_{ij})_\alpha = r_{ij}$ while $r_{ij1} = r_{ij2}$. $(r_{ij})_\alpha$ gets the lower limit or upper limit of the interval while $\alpha(0 \le \alpha \le 1)$ equals 0 or 1.

$$\sum_{i=1}^{m} w_i = 1$$

To note is, in the weight fuzzy process, normalization should be combined. The fuzzy preference weights are mid-value normalized.

E. DEA Traditional Model

Suppose n DMUs, and each DMU has m input indexes, s output indexes, i.e.:

$x_j = (x_{1j}, x_{2j}, ..., x_{mj})^T$, $y_j = (y_{1j}, y_{2j}, ..., y_{sj},)^T$ are accordingly input or output index vector of the $j(j = 1,2,...,n)$ DMU; and

$w = (w_1, w_2, ...w_m)^T$,$v = (v_1, v_2, ..., v_s)^T$ are accordingly input or output weight vector, E_k is fuzzy relative evaluation of the k DMU. The traditional DEA model is [12]:

[Model 1]

$$\max v^T y_k = E_k$$

$$s.t. \begin{cases} w^T x_k = 1 \\ v^T y_j - w^T x_j \le 0, (j = 1,2,...,n) \\ w \ge 0 \\ v \ge 0 \end{cases}$$

In model 1, DMU requires input index x_j and output index y_j $(j = 1,2,...,n)$. Model 1 is not applicable with only input or output. The evaluation does not take into account possible preference between indexes, even fuzzy preference information. The improved model of fuzzy DEA is required to solve the problem.

F. Fuzzy Preference Weight

Linear inequality incomplete weight information can be described as the following basic forms [13]:

$$\{w_i \ge w_j\} \quad (1)$$

$$\{a_i \le w_i \le a_i + \varepsilon_i\} \quad (2)$$

Wherein, a_i and ε_i are nonnegative constant; order weight information in (1) and interval number weight information in (2), they both are the most common incomplete weight information. The remaining forms are always based on them. In general, these integrated constraints can be expressed as a linear inequality $\boldsymbol{Cw} \leqslant \boldsymbol{b}$ [12].

Furthermore, $\boldsymbol{Cw} \leqslant \boldsymbol{b_f}$ is introduce where $\boldsymbol{b_f}$ contains fuzzy preference weight information.

In this way, the fuzzy information expression $\boldsymbol{Cw} \leqslant \boldsymbol{b_f}$ constructs a window on the regulation of weights information, called fuzzy window. Subsequently, cut sets are applied in model simulation method for different confidence levels α , and $(b_f)_\alpha$ describes the fuzzy opening degree of the fuzzy windows. So that more progressive details can be analyzed in evaluation results of the model.

G. Improved DEA Model

Each program to be evaluated is taken as a DMU. The constraint based on fuzzy window ($\boldsymbol{Cw} \leqslant \boldsymbol{b_f}$) is formed with all its fuzzy output indexes and linear expressions. Then the improved fuzzy DEA model with fuzzy preference information for output only is shown in the following:

[Model 2]

$$\max \sum_{i=1}^{m} w_i r_{ik} = E_k$$

$$s.t. \begin{cases} \sum_{i=1}^{m} w_i r_{ij} \le 1, (j = 1,2,...,n) \\ Cw \le b_f \\ w_i \ge 0 \end{cases}$$

Wherein, E_k is the evaluation index of the program k;

$w = (w_1, w_2, ...w_m)^T$ is one of weight valuables of indexes;

C is p × m matrix, here p is constrains number of weights;

$b_f = (b_1, b_2, ...b_n)^T$ is a fuzzy vector including fuzzy preference weight information of indexes.

In the model, for a given incomplete information of weights, based on quantify results of fuzzy preference weight information, the parameters of $\boldsymbol{C}$ and $\boldsymbol{b_f}$ are set appropriately and fuzzy windows constrain $\boldsymbol{Cw} \leqslant \boldsymbol{b_f}$ is built. The model solves the practical evaluation with fuzzy preference weight information

H. Solving Algorithm

Model 2 is of a fuzzy mathematical planning type, and it is to be transformed to a deterministic model for solution. All the fuzzy numbers in model 2 are expressed as cut sets interval of $\alpha(0 \le \alpha \le 1)$.

[Model 3]

$$\max \sum_{i=1}^{m} w_i (r_{ik})_\alpha = (E_k)_\alpha$$

$$s.t. \begin{cases} \sum_{i=1}^{m} w_i (r_{ij})_\alpha \le 1, (j = 1,2,...,n) \\ Cw \le (b_f)_\alpha \\ w_i \ge 0 \end{cases}$$

Therefore, by the cut sets of fuzzy numbers, the fuzzy evaluation model is transformed to determinate planning model with α cut set. The evaluation with different confidence level $\alpha(0 \le \alpha \le 1)$ can be obtained and known well.

III. RESULT AND DISSCUSION

Five automatic testing softwares, named program 1 to 5 respectively, are to be evaluated. The subjective evaluation of the indexes are shown in Table II.

TABLE II. EVALUATION INDEXES AND SUBJECTIVE VALUES

Index	Prog.1 (x_1)	Prog.2 (x_2)	Prog.3 (x_3)	Prog.4 (x_4)	Prog.5 (x_5)
Functionality (I1)	very poor	middle	good	poor	excellent
Reliability (I2)	very poor	middle	good	poor	excellent
Usability (I3)	middle	middle	excellent	good	excellent
Efficiency (I4)	middle	middle	excellent	good	excellent
Mantainability (I5)	very poor	middle	poor	excellent	middle
Portability (I6)	very poor	middle	poor	excellent	middle

Transform the subjective indicators values to fuzzy numbers, as shown in Table III.

TABLE III. EVALUATION INDEXES AND SUBJECTIVE VALUES

Index	Sol.1 (x_1)	Sol.2 (x_2)	Sol.3 (x_3)	Sol.4 (x_4)	Sol.5 (x_5)
I1	(0, 0.3)	(0.4, 0.7)	(0.6, 0.9)	(0.2, 0.5)	(0.8, 1.0)
I2	(0, 0.3)	(0.4, 0.7)	(0.6, 0.9)	(0.2, 0.5)	(0.8, 1.0)
I3	(0.4, 0.7)	(0.4, 0.7)	(0.8, 1.0)	(0.6, 0.9)	(0.8, 1.0)
I4	(0.4, 0.7)	(0.4, 0.7)	(0.8, 1.0)	(0.6, 0.9)	(0.8, 1.0)
I5	(0, 0.3)	(0.4, 0.7)	(0.2, 0.5)	(0.8, 1.0)	(0.4, 0.7)
I6	(0, 0.3)	(0.4, 0.7)	(0.2, 0.5)	(0.8, 1.0)	(0.4, 0.7)

The direct subjective expectations of each index are shown in Table VI with mid-value normalization of fuzzy numbers.

TABLE IV. DIRECT SUBJECTIVE FUZZY PREFERENCE OF INDEXES

Index	Fuzzy preference	Fuzzy interval	Normalization
I1	concerned most	(0.8, 1.0)	(0.25, 0.3125)
I2	concerned more	(0.6, 0.9)	(0.188, 0.281)
I3	tightly concerned	(0.2, 0.5)	(0.0625, 0.156)
I4	tightly concerned	(0.2, 0.5)	(0.0625, 0.156)
I5	concerned more	(0.6, 0.9)	(0.188, 0.281)
I6	tightly concerned	(0.2, 0.5)	(0.0625, 0.156)

The fuzzy intervals of weights are expressed as form of $a_i \leqslant w_i \leqslant b_i$, and used to construct fuzzy window $\boldsymbol{Cw} \leqslant \boldsymbol{b_f}$

$$\sum_{i=1}^{m} w_i = 1$$

with normalization of $_{i=1}$. Then the instance of model 3 is built. Moreover, the evaluation interval results (x_{ml}, x_{mu}) (m=1,2,…5) can be obtained according to different α level by combining data in Table III.

TABLE V. EVALUATION RESULT BASED ON THE DIRECT SUBJECTIVE PREFERENCE FUZZY WINDOW

α值	0	0.2	0.4	0.6	0.8	1.0
x_{1l}	0.063	0.124	0.185	0.246	0.307	0.368
x_{1u}	0.104	0.171	0.238	0.305	0.373	0.440
x_{2l}	0.407	0.468	0.529	0.590	0.651	0.712
x_{2u}	0.448	0.515	0.582	0.649	0.716	0.783
x_{3l}	0.516	0.574	0.632	0.690	0.748	0.806
x_{3u}	0.578	0.640	0.701	0.763	0.825	0.887
x_{4l}	0.454	0.509	0.563	0.618	0.673	0.728
x_{4u}	0.546	0.606	0.666	0.726	0.786	0.845
x_{5l}	0.688	0.735	0.782	0.829	0.876	0.923
x_{5u}	0.749	0.801	0.853	0.905	0.957	1.000

According to data in Table V, the evaluation membership curves of each program are shown in Figure 1 and Figure 2.

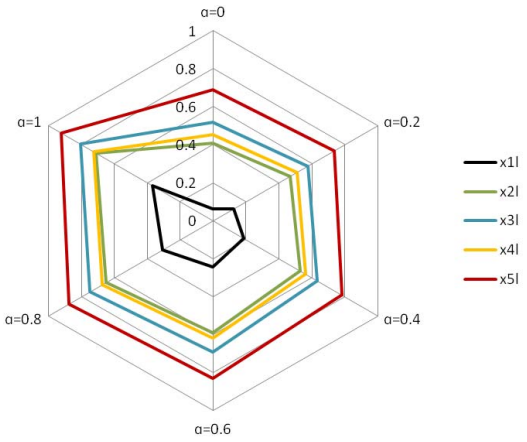


Figure 1. Lower limits evaluation membership curves based on range fuzzy window

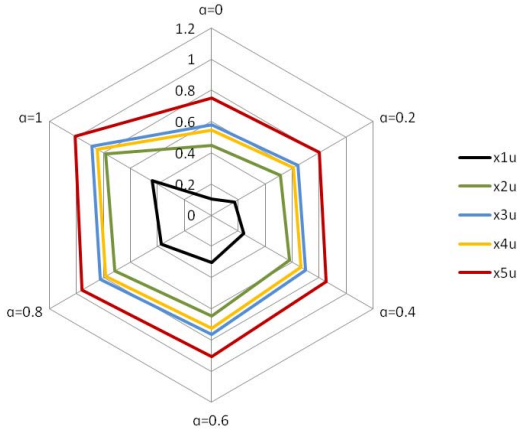


Figure 2. Upper limits s evaluation membership curves based on range fuzzy window

It can be concluded that the evaluation results keep the consistence for different levels of α .

Therefore, based on the improved DEA model and its solution, evaluation and decision making can be done with both fuzzy index and fuzzy preference information. The comprehensive evaluation is achieved at different confidence levels of α and fuzzy window $\boldsymbol{Cw} \leqslant \boldsymbol{b}_f$.

IV. ARGUMENTATION AND CONCLUSION

Based on the DEA model, an improved evaluation model is introduced for the fuzzy index value and fuzzy preference weight information. The improved model is applied in automatic testing software programs evaluation. Firstly, the subjective index value is transformed to fuzzy numbers. Secondly, the subjective preference weights are constructed as fuzzy window constrain. Finally, the improved fuzzy DEA model is established to obtain the comprehensive evaluation of automatic testing software programs. A corresponding solving algorithm for the evaluation model is practiced.

The evaluation results of automatic testing software programs keep consistence for different confidence levels with direct subjective weight preference information. For other forms of fuzzy index value and fuzzy weight information, the solving process is similar.

ACKNOWLEDGMENT

The authors would like to thank to the project of Ministry of Education of PRC (no.11YJC630106).

REFERENCES

[1] Tu Ling, Zhou Yan-hui, Zhang Wei-qun, Zhou Ya-zhou, “Fuzzy Logic Based Metric in Software Testing,” Comuter Science, vol.36, no.7, pp. 141-144, Jul. 2009.

[2] Yang Ming-shun, Li Yan, Lin Zhi-hang, “Combinatorial decision-making on product outline schemes evaluation & optimum selection,” Computer Integrated Manufacturing Systems, vol.12, no.4, pp. 540-545, Apr. 2006.

[3] Sun Hua-li, Wang Ji-qiang, Wen Zheng-zhong, “A Study on Green Design and its Fuzzy Assessment Method,” Machanical Science and Technology, vol.22, no.5, pp. 699-704, May 2003.

[4] WANG J., “Ranking engineering design concepts using a fuzzy out ranking preference model,” Fuzzy Sets and Systems, vol.119, no.1, pp. 161-170, Jan. 2001.

[5] KHAN F I, SADIQ R, HUSAIN T. GreenPro-I, “A risk-based life cycle assessment and decision-making methodology for process plant design,” Environmental Modeling and Software, vol.17, no.8, pp. 669-692, Aug. 2002.

[6] LE TENO J F, MARESCHAL B., “An interval version of PROMETHEE for the comparison of building products' design with ill-defined data on environmental quality,” European Journal of Operational Research, vol.109, no.2, pp. 522-529, Feb. 1998.

[7] GELDERMANN J, SPENGLER T, RENTZ O, “Fuzzy out-ranking for environmental assessment. Case study: iron and steel making industry,” Fuzzy Sets and Systems, vol.115, no.1, pp. 45-65, Jan. 2000.

[8] CHIOU H K, TZENG G H., “Fuzzy multiple-criteria decision-making approach for industrial green engineering,” Environmental Management, vol.30, no.6, pp. 816-830, Jun. 2002.

[9] Liu Ying-ping, Lin Zhi-gui, Gao Xin-ling, Shen Zu-yi, “Study on Greenness Evaluation Method with Incomplete Information on Weights of Multilevel Indices,” China Mechanical Engineering, vol.17, no.1, pp. 29-33, Jan. 2006.

[10] Liu Ying-ping, Gao Xin-ling, Shen Zu-yi, “Product design schemes evaluation based on fuzzy DEA,” Computer Integrated Manufacturing Systems, vol.13, no.11, pp. 2099-2104, Nov. 2007.

[11] China Quality Inspection press fourth editing dept., National standard of computer software engineering (2nd Edition), CN:Chinese Standards Press, pp. 52-232, 2011.

[12] LERTWORASIRIKUL S, FANG S C, JOINES J A, et al.，”Fuzzy data envelopment analysis (DEA)-a possibility approach,” Fuzzy Sets and Systems, vol.139, no.2, pp. 379-394, Feb. 2003.

[13] Wang Jian-qiang, “Study on hierarchical discrimination approach of multi-criteria classification with incomplete information,” Control and Decision, vol.19, no.11, pp. 1237-1245, Nov. 2004.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Gender difference in the use of hospitalization services in rural China ——evidence from Sichuan province

Ye Shaoxia
School of information management
Wuhan University
Wuhan China
wsshaoxia@126.com

Yin Cong
School of information management
Wuhan university
Wuhan China
cyinwhu@126.com

***Abstract*—Gender differences in health use have been documented widely while results differ between studies and countries. We try to explore the differences between gender in rural China employing 669，000 hospitalization data of rural counties of Sichuan province in 2013. Descriptive statistics were employed to calculate the sums, means and proportions of the admission , length of stay and expenditures of hospitalization. The Chi-square test was used to calculate differences between proportions and the t test was used to test differences between means. Our results suggested that women had more hospital utilization than men while men had higher duration and expenditures of hospitalization. More female were admitted to hospital than male across all age. 25-34year is a key age range where female/male admission ratio was highest (2.1:1)and male had a longest duration(11.6days) in all life. Male had higher expenses of hospital than female in all age except infancy and the gap in expenses between genders is rare before adult. The differences of gender in health use does exist and policies and economic support should be taken to rural residents.**

Keywords- gender differences; hospitalization; rural China;hospital expense.

I. INTRODUCTION

Gender as a social category usually indicates the specific role, responsibility and social status in society. Men are in the dominant position in almost every field of social life such as in education, employment, political and economic empowerment[1]. Although today extreme sexist views about women are rare, and women even surpass men in educational attainment[2], but women still have limited access to some occupation areas and organizations pervasively. Inequality in access to these resources may produce stresses and strains over time, resulting the decline in the overall quality of life of women. Although researchers argued that female have advantage in some health outcomes ,such as a longer life expectancy and lower mortality ,due to the biological and behavioral factors [3], evidences also obviously suggest that women have been brought to greater sickness and hospital use than men because of the universal inequalities in the social life[4].

Many researches suggested that the mortality rates of men are higher than those of women, while women's morbidity rates are higher than those of men[5] . Women have been consistently reported to make greater use of healthcare services and have higher expenditures than men, even when reproductive care is excluded[6]. Women have a greater chance of suffering from chronic diseases while men tend to suffer from life-threatening diseases[7]. The explanations of the differences have been constantly explored ,and physiological structure and social characteristics both may lead to the differences in health and medical services use including the early life conditions, unhealthy behaviors, healthcare utilization, health knowledge, reproductive function and the expression of genes [8].

The results of studies in China on gender differences in medical services utilization are inconsistent with foreign or local researches. The study on Zhuhai, a high developing economic zone in China, applied hospitalization data of the city to analyze gender differences in health use showed that men had a significantly higher admission, longer duration of hospitalization and higher expenditures than women[9]. Research in Chinese poor rural area on sampling survey showed that female got more medical services in most age group with females having higher two weeks prevalence and higher prevalence of chronic disease than males[10]. Ofra Anson found that the medical use among women aged 25-44 year was higher than that of men and reached peak after 65 years of age[11].

Empirical studies on gender difference in health care in China are rare and there are some shortcomings. Most of the researches collected data from questionnaires only[11][12].The study involved in comprehensive and objective data is often confined to a hospital or a small area with limited representation[9]. It is rather important to discover the gender difference of medical utilization on full volume data of rural China where economic development is relatively backward with the access to medical resources more difficult[13] and traditional concept ingrained. We try to figure out whether there is a gap between men and women in medical services, and what's the difference between our results and those of previous studies. New rural cooperative medical care (NCMS)is the medical insurance covering more than 98% of the Chinese rural population which comprehensive records the medical services of rural residents. We applied NCMS 669,000 inpatient data to analyze gender difference in inpatients in rural China. The

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.118

findings has important implications for further policy design and promotion of equitable access to health care.

II. METHODS

A. Data source

The data used in our study were obtained from New rural cooperative medical care system containing the whole inpatients records of 15 counties in Sichuan province in year 2013. Sichuan province had the third population in China with the province's total GDP ranking 9 and per capita income ranking 25 in 2013. Located in the mid-west of China, to some extent, it can reveal the characteristics of Chinese rural. The 15 counties chosen belong to 10 cities and per capita income of farmers ranged from 5000 to 11800yuan, population ranged from 3.2 to 97 million so the sample s were random enough to ensure the representative of the data. The total participation in NCMS of 15 counties is 646,6 million, with an admission of 66.96 million to hospital in one year (10.4%). Distribution of hospitalization rate in different counties varied in 6.3%-20.3%.

B. Study variable

In order to analyze the utilization of hospital services, specific variables are employed including gender, age, hospital grade, surgical therapies, length of hospitalization, hospitalization expenditure, compensation fee and the category of disease. The age is divided into 8 segments with the infancy(0-4 years) and elderly(>65years)as two group and the rest was cut to groups every ten years. Hospital level is divided and optimized into 5 grades according the "Hospital classification standard" :grade 1 to 5 respectively refers to the village clinic, township hospital, county hospital, municipal hospital, and hospital above municipal hospital. The hospital's comprehensive ability increase with the level of hospital rise. Length of hospitalization is divided into 5 groups: within ten days, one month, two months, 100 days and 100 days above where the one month group refers to the length between ten days and 30 days. Disease category is defined according the "tenth edition standard of the international classification of disease".

C. Statistical analysis

Hospitalization data were stored in Oracle Database with statistical analysis performed with IBM SPSS version 19.0 and Microsoft Office Excel 2007. Descriptive statistical analysis were applied to account the sum, mean and proportion. Chi-square test was used to calculate differences between proportions and t test was used to test differences between means. Additionally, during analysis, the total medical expense per inpatient was transformed into natural logarithms of the observation value to address the positive skew of the expenditure data. Results with a 2-sided p value <0.05 were considered statistically significant

III. RESULTS

The total admission of hospitalization in 2013 was 669,000. The distribution of the special attributes was shown by Table 1 to compare the age, length of hospital , hospital grade and surgical therapies between female and male. Comparison of the means of age, length of stay and cost of hospitalization was shown in Table 2.

The total number of female inpatients was higher than that of males(364011, 305631) in a 1.19:1.0 ratio. The average age of female was slightly lower than male(48.24，48.55). More female were admitted to hospital than male across all age groups except in infancy and young (age<15) where the male/female ratios were both 1:0.74 . The admission of female was higher than that of male obviously in adult period and the gap reached the peak in aged 25-34 in a ratio 2.14:1.0. Overall, men had a significantly longer duration of hospitalization than women(mean duration,10.65 vs. 9.76 days, respectively; p=0.00). 95.8% of hospitalized patients were discharged within a month, and only 0.5% patients were hospitalized for more than one hundred days. The proportion of men increase with the rise of the length of hospitalization where male/female ratios in group within ten days and above hundred days were respectively1:1.25 and 1:0.82. Regarding the grade of hospital, patients were mainly distributed in the grade 2 and grade 3 hospitals (46.4%, 41.1%). With the increase of hospital level, the number of hospital admission decreased for both female and male. And the admission of female was outstripping men in all level of hospital. Furthermore, a larger proportion of hospitalized female underwent surgery compared to male(19.1%vs.22.9%,respectively; p = 0.00).

TABLE I. MAIN CHARACTERISTICS OF ALL PATIENTS ADMITTED TO HOSPITAL BY GENDER

Variable		Male (305631)	Female (364011)	Female/ Male (1.19)
Age range (year)	<5	20789	15331	0.74
	5-14	25962	19470	0.75
	15-24	13293	25742	1.94
	25-34	15005	32110	2.14
	35-44	36543	51706	1.41
	45-54	40101	54403	1.36
	55-64	61568	69376	1.13
	>65	92370	95873	1.04
Length of stay (day)	ten days	205165	257233	1.25
	one month	85655	93148	1.09
	two month	9663	9180	0.95
	100 days	3323	2956	0.89
	above 100 days	1825	1494	0.82
Grade of hospital	2	140514	170076	1.21
	3	126989	148242	1.17
	4	30221	35100	1.16
	5	7907	10593	1.34
Surgical therapies	no	247250	280494	1.13
	yes	58381	83517	1.43

Regarding the medical expenses, gender differences with statistical significance were observed. The average expenses of hospitalization of male were significantly higher than that of female (3125vs.2922yuan,p=0.00) with a gap of 200yuan, and reimbursements for expenses of male were as well

significantly higher than that of female (1877vs.1730yuan,p=0.00) with a gap of 140yuan. The proportion of compensation for male and female was respectively 65.2%, 64.6%.

TABLE II. MEANS OF VARIABLES BY GENDER IN HOSPITAL

	Male	Female	Gender difference	P value
Age, year	48.55	48.24	0.60%	0.00
length of stay, day	10.65	9.76	8.40%	0.00
expenditure of hospital, yuan	3126	2923	6.50%	0.00
Compensation fee, yuan	1878	1731	7.80%	0.00

For more detailed comparison, figure 1 and figure 2 showed the distribution of the differences of length of stay and hospitalization expenses between male and female in varied age groups. With regard to the length of stay ,male had a longer duration across any age group and the gap grew bigger while age was closer to 30year which reached the peak in age 25-44 year. Even in age group 25-34year when the admission of women was highest ,the average length of hospitalization of men was higher than women in a gap of 2.2 days. There was almost no difference in length of hospitalization between male and female in age group <4year and >65year. As to the expenditures of hospitalization, the costs of hospitalization of male were higher than that of female across all age group except for infancy, and that reached highest in age 45-64year,and that when the differences between gender become biggest. Before middle age, the gap of the hospitalization expenses was not obvious, especially in the infancy(age<5), the cost of hospitalization of female is even slightly higher than male (1826vs. 1793yuan).

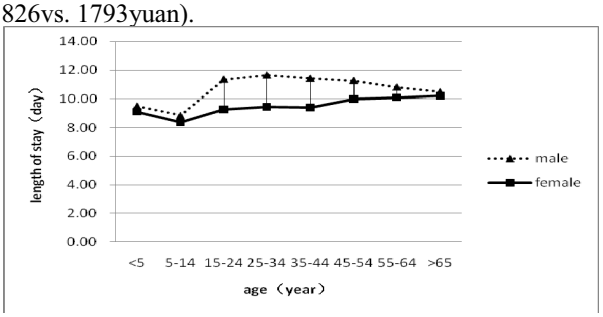


Figure 1. Length of stay by gender, by age

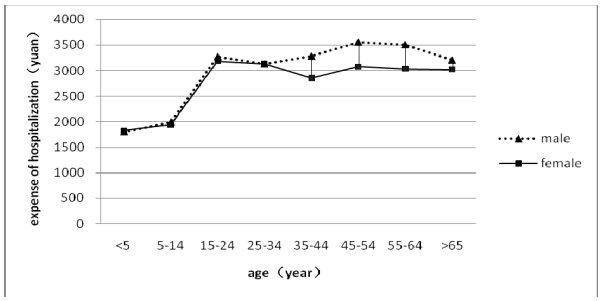


Figure 2. Expenses of hospitalization by gender ,by age

We tried to figure out what categories of disease lead to the big difference of admission and how about the differences of their length and expenditure of hospitalization. We selected six kinds of diseases three of which refer to the diseases in highest ratio of female/male in admission and three of which refer to the lowest ratio diseases thus we knew the diseases favor female and male respectively. The three diseases with a outstanding prevalence in male versus female were diseases of the skin and subcutaneous tissue(L00-L99), injury, poisoning and certain other consequences of external causes(S00-T98) ,and external causes of morbidity and mortality(V01-Y98). Vice versa diseases favored women were diseases of the ear and mastoid process(H60-H95), diseases of the genitourinary system(N00-N99), and diseases of the blood（D50-D89) when excluding the reproductive care which was specific for women. The results suggested that gender usually suffer a higher incidence disease in a younger age(Table 3). Males' average age in diseases favor them was significantly lower than that of females. Regarding the length of hospital stay and hospitalization expenses, men had a generally higher value than women in each of the six diseases, meaning men were more severe than female in any diseases whether in the high or low incidence diseases. Furthermore the average expenses of diseased favor men were significantly higher than those favor women ,which means men were prone to critical diseases.

TABLE III. AVERAGE OF VARIABLES IN HIGH INCIDENCE DISEASES BY GENDER

Diseases	**Age,year**		**Length of stay,day**		**Expenses of hospital,yuan**	
	Male	Female	Male	Female	Male	Female
H60-H95*	50.46	54.09	8.63	7.44	1733	1428
N00-N99*	47.11	40.81	9.07	8.77	2974	2578
D50-D89*	54.25	48.43	10.89	9.93	3875	3482
V01-Y98^	40.92	47.7	20.31	18.98	5859	5560
L00-L99^	44.06	44.52	10.12	10.05	1964	2025
S00-T98^	43.1	50.01	14.57	14.09	4717	4607
Total	48.55	48.24	10.65	9.76	3126	2923

Disease was represented by code,*refer diseases favor female, ^refer diseases favor male.

IV. DISCUSSION

In this study, we analyzed the gender differences in the use of inpatient services for rural residents in 15 counties, and the results can reflect the status of medical use in rural areas of China. Within the range of the selected characteristic, there were significant differences in the utilization of inpatient services between men and women. Our main conclusion was that female had more utilization of hospitalization than male while male's illness severity was

higher than that female. The hospital admission of female was significant higher than male across all age except infancy with the gap reach the top in child-bearing age in a female/male ratio 2.14:1. The findings are consistent with previous studies. Many studies suggested that female had a higher morbidity rate than male in all life only except infancy age with the mortality and the morbidity reaching the highest in the reproductive age[11]. Even adjusting for pregnancy female had a higher utilization of medical services than male[14]. Some researchers found that the morbidity rate of male tend to the top in age 40-69 [13]which was also demonstrated in our study. Men had a longer length of hospitalization than women in all age with the difference expand in middle age [11]which also verified the previous study. We found that male had longer duration in hospital than male across all age, and the male/female ratio in duration over 100days even reached 1:0.82. As expected, the length of the hospital stay is a key factor affecting the cost of hospitalization[15]，thus male had a higher expenditure than female overall and the growth trend of expenditures was consistent with that of the length of hospital stay . Age group in15-24year referred a high morbidity in female while a high expenses in male, which meant male in middle age had more severe diseases although female faced the reproductive disease. We found that male's expenses were increasing along lifetime and reached the highest in 45-64year while female's expenses changed flat in adult. This finding was consistent with some studies suggesting that male medical expenditure is higher than that of female and especially high after middle age[16][17].

Previous researches and traditional perspective are of the view that women are vulnerable groups having less access to health services but in fact the inpatient utilization of female is high while the length of hospital stay and costs of female is rather lower than men. As the data analyzed in our study is the NCMS inpatient data in rural China which represents a group of poorer population with relatively slow economic development. There are some possible explanations for our results: First of all, the rural men compared to women engaged in high intensity work easily cause the serious disease, which is verified by the high incidence of injury and external causes of morbidity and mortality in our analysis. Second, women have a high probability of occurrence of chronic diseases which means although the average length of stay and expenses of hospitalization are low ,they are more prone to high frequency of inpatient and outpatient. Third, men tend to seek for treatment until the illness is serious when they can hardly stand it, resulting in the higher cost and longer stay of hospitalization. Finally, female may prefer to cheap drugs and short time hospitalization since the low economic status in family. Deeper and crucial reasons for our results need further demonstration.

There are some limitations in our study. Although the data of 15counties represents a part of China, it can hardly refer the whole situation of China. At the same time, we only compared the differences in hospitalization services with no consideration of outpatient services and self-medication. Gender differences in the use of health care do occur in rural China and we wish to balance the differences between gender through the policies and social efforts. In the future on-going studies on the same topic should be taken in bigger settings covering more services which will have important significance for China's rural health and health equity between gender.

REFERENCES

[1] Li CX, Wu ZC, Xu L, Gao J. Gender differences in medical expenditure in China. Chinese Health Eco Magazine ,2006, 25:46–48.

[2] Ryan,C.L.& Siebens J. Educational Attainment in the United States : 2009. Retrieved from http://www.census.gov/prod/2012pubs/p20-566.pdf.

[3] Preston, Samuel H. and Haidong Wang. Sex mortality differences in the United States: The role of cohort smoking patterns. Demography,2006,43(4): 631-646.

[4] Cummings, J.L. and P.B. Jackson. Race, Gender, and SES Disparities in Self-Assessed Health, 1974-2004.*Research on Aging*, 2008,30(2): 137-168.

[5] Arber, S., & Cooper, H.. Gender and inequalities in health across the life course. In E. Annandale, & K. Hunt (Eds.), 1999 , 123–149.

[6] Cylus J, Hartman M, Washington B, et al. Pronounced gender and age differences are evident in personal health care spending per person. Health Aff (Millwood) ,2011,30:153-60.

[7] Rieker, Patricia P. and Chloe E. Bird. "Sociological Explanations of Gender Differences in Mental and Physical Health".2000,Pp. 98–113 in Handbook of Medical Sociology edited by C. E. Bird, P. Conrad, and A. M. Fremont. Upper Saddle River, NJ: Prentice-Hall.

[8] Vera Regitz-Zagrosek. Sex and gender differences in health. European molecular biology organization, 2012, 13 :596-603.

[9] Yan Song, Ying Bian .Gender differences in the use of health care in China: cross-sectional analysis. International Journal for Equity in Health, January 2014, 13:8.

[10] Huang Chengli. The Gender Differences between Medical Services Utilization and Medical Expenditure in Chinese Poor Rural Area. Market and Demographic Analysis, 2003,2:37-43.

[11] Ofra Anson, Shifang Sun. Gender and health in rural China: evidence from HeBei province. Social Science & Medicine, 2002,55:1039–1054.

[12] Hsiu-Ju Chang, Yuen-Liang Lai, Chia-Ming Chang, Ching-Chiu Kao, Meei-Ling Shyu, Ming-Been Lee. Gender and Age Differences Among Youth, in Utilization of Mental Health Services in the Year Preceding Suicide. Community Ment Health J ,2012,48:771–780 .

[13] ZongFu Mao, Bei Wu. Urban–rural, age and gender differences in health behaviours in the Chinese population: findings from a survey in Hubei, China. Public Health ,2007, 121:761–764.

[14] Cylus J, Hartman M, Washington B, et al. Pronounced gender and age differences are evident in personal health care spending per person. Health Aff (Millwood) ,2011,30:153e60.

[15] Zhao Fen, Liu Jin-lin, Wu Jing—xian, Jing Peng-pen. An Analysis on Influencing Factors of Hospitalization Expenditure Under the Single DRGs Mode. Journal of Northwest University (Philosophy and Social Sciences Edition),2014,44(3):170-176.

[16] Li Chuxiang, Wu Zhuochun, Xu Ling, et al. Gender Differences in Medical Expenditure in China .Chinese Health Economics, 2006, 25(2) :46- 48.

[17] Ynag Jiannan , et al .Comparative Analysis of the Medical Cost of the Male or the Female Inpatients Hospitalized in General Hospital. Journal of Mathematical Medicine, 2011,24(2):187-189.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

The empirical analysis on the influential factors of urbanization in Hubei province based on the panel data

Yu Qing, Wu Xiaoyuan,Yuan Meng,Xiong Qian, Jin Shengping*
Department of Statistics, Wuhan University of Technology, Wuhan, China
*corresponding author: spjin@whut.edu.cn

***Abstract*—We used 12 indicators of 11 important cities from 6 years to set up panel-data model. By using Eviews, we firstly conducted Unit root test and selected variables according to stationarity. Secondly, to identify co-integration relationship we carried out co-integration test. Thirdly, 2 variables--number of units employed in the end of the year and Social security employment expenses are selected by doing Haus- man test and F test. Finally we applied the fixed coefficient model to study influential factors of urbanization in Hubei province empirically. We have got the conclusions: In the urbanization level to number of units employed in the end of the year model, the employment situations in highly urbanized areas are more sensitive. In the urbanization level to Social security employment expenses model, the effect of Social security employment expenses on the urbanization mainly depends on the level of urbanization.**

Key words: urbanization rate; panel-data, unit root test, co-integration test, the fixed-coefficient model

I. INTRODUCTION

Hu Jintao's report at 18th Party Congress said: 'We should keep to the Chinese-style path of carrying out industrialization in a new way and advancing IT application, urbanization and agricultural modernization. We should promote integration of IT application and industrialization, interaction between industrialization and urbanization, and coordination between urbanization and agricultural modernization, thus promoting harmonized development of industrialization, IT application, urbanization and agricultural modernization.' This shows that, the new urbanization has rich connotation and integrated characteristics, therefore we define it not from one perspective or one level, but from both commonness and individuality, generality and particularity. Urbanization on the one hand is the process of population from rural to urban migration accumulation, on the other hand shows the change of regional landscape, the change of industrial structure, the change of way to product and live, also Urbanization is an integrated unified process of population, geographical, social and economic organization , production and life style from the traditional backward society to modern society, so it reflects a country or a region's economic and social development and progress.

Domestic scholars have done a lot of researches in population, economic, social security and other aspects of the influence of urbanization. Most of them studied single factor of the urbanization process. However, we establish Panel data model to explore the effects on the process of urbanization from population, economy, life level and social security four aspects. We find the main factors, leading factors and lag factors which are influencing new urbanization from theory and practice. The conclusions are significant: providing credible proposal for new development strategy of urbanization in Hubei province; under the limited resource and environmental constraints, helping the relevant departments guide each department making the most powerful contribution for the deployment of new urbanization,; giving monitoring and adjustment suggestions on comprehensive planning of infrastructure, including regional urban system planning and urban planning and so on.

II. INDEX SELECTION AND DATA SOURCES

A. *Selection of explanatory variables*

The select of explanatory variable and explained variable should not only consider the representative, also consider the availability of data. Therefore, we select the representative population, economic development, social security and employment, people's livelihood four aspects of 12 explanatory variable for analysis.

The population[1]. Considering the relationship between urban population and the land demand, the regional population flow influence on the rate of urbanization, we select regional population density as the index of population distribution.

Economic development[2]. Referring to relevant data, we select GDP per capita, The primary, second and third industry gross domestic product, urban fixed-asset investment, reception of domestic tourism income these indicators to measure economic development situation in each region and then explore their relation with regional urbanization.

The social security[3]. We select the urban and rural community affairs expenses, Cultural and education section operating expense, social security and employment expenses as indicators to analyze their effects on urbanization in different regions.

People's livelihood. We select the consumer price index as the measure of people's life.

B. *The determination of explanatory variable*

The explanatory variable in our thesis is urbanization rate of different area. We usually use the urbanization rate

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.119

based on permanent population to reflect the level of urbanization of a country or an area. Our thesis use the urbanization rate based on household registration to describe the level of urbanization in china. On one hand, it can show the current situation of urbanization in china and its influence on the development of urbanization, on the other hand, it also can describe the space and the direction in future that the urbanization process of china towards.

Due to the collection of indicators in statistical yearbook changing every year and data missing problem, we finally choose the index data of 2008-2013 in 11 municipal city in hubei province (wuhan, huangshi, shiyan, yichang, xiangyang, ezhou, jingmen, xiaogan, jingzhou, huanggang and xianning)after repeatedly weigh. All the data are from 'statistical yearbook of Hubei province 'each year and the 'the China's urbanization rate survey report' White Paper.

III. MODEL BUILDING

Based on the panel data which contain 12 indicators of 11 important cities of Hubei provinces from 2008 to 2013, firstly we make unit root text and cointegration test to make sure whether there is long-run equilibrium relationship between each indicators and the urbanization rate[4].

A.The results of unit root test

According to the results of unit root test, we can know that except A， the absolute value of all explained variables' ADF statistics are higher than the critical level of 0.05, the original series all accept the null hypothesis under the significanc level of 0.05, so those time series except A are nonstationary series. After the first order difference, there are only $LOG(B)$、LOG（G）、LOG（J）、$LOG(K)$ and $LOG(M)$ rej ect the null hypothesis under the significanc level of 0.05. That means those first difference series are stationary series under the significanc level of 0.05, so those five series are $I(1)$.Therefore, we choose B, G, J, K and M to do the cointegration test.

B.The result of Pedroni panel cointegration test

TABLE I. The test results of the index K (Pedroni)

	statistic	P value
Panel v	-0.003296	0.5013
Panel rho	-1.48551	0.0687
Panel PP	-10.43732	0.0000
Panel ADF	-10.64113	0.0000
Group rho	0.273476	0.6078
Group PP	-13.83939	0.0000
Group ADF	-15.32788	0.0000

From Tab. I, we can see that among the cointegration test results of index K, there are four P value is less than 0.05, which we claim that there is a long-term equilibrium relationship between index K and urbanization rate, and it can move on to do the next step regression analysis. The other three indexes' test results are also contains four P value less than 0.05, so that they can take the next step regression analysis.

C.Choosing the styles of model

TABLE II. The results of H test and F test

Dependent Variables	P value of H test	F_1	F_2
$\log(B)$	0.356	1.09	11.6
$\log(G)$	0.197	59.6	11.3
$\log(J)$	0.0578	15.9	76
$\log(K)$	0.5017	0.44	16.8

Tab. II about H test results shows that four dependent variables' p values are greater than 0.05, there are fixed effects between dependent and independent variables. The F test result of Per Capita GDP (yuan) $LOG(B)$ shows that $F_2 = 11.6 > F(20,44) = 1.813898$,it is reasonable to refuse H_2 ,and because $F_1 = 1.09 > F(10,44) = 2.053901$,it's reasonable to accept H_1 , so we adopt the fixed effect and fixed influence model; The F test result of the unit number of on-the-job at the end of a year(ten thousand people) $LOG(G)$ shows that $F_1 = 59.6 > F(20,44)$,it is reasonable to refuse H_1 ,and because $F_2 = 15.3 > F(10,44)$,it's reasonable to refuse H_2 , so we adopt the fixed effect and variable influence model. The F test result of social security and workers employment spending (one hundred million yuan) $LOG(J)$ shows that $F_1 = 15.9 > F(20,44)$,it is reasonable to refuse H_1 ,and because $F_2 = 76 > F(10,44)$,it's reasonable to refuse H_2 , so we adopt the fixed effect and variable influence model. The F test result of the consumer price index (Last year was 100) $LOG(K)$ shows that $F_2 = 16.8 > F(20,44)$, it is reasonable to refuse H_2 , and because $F_1 = 0.44 < F(10,44)$, it's reasonable to accept H_1 , so we adopt the fixed effect and fixed influence model.

D. Modeling

Following, we begin to use the Eviews to estimate the model parameter and get the final result, which can show the influence of each indicators on the urbanization of areas.

Now, we can establish a fixed effect variable model, the model specific parameter settings are as following:

$$y_{it} = \alpha_0 + \alpha_i + x_{it}\beta_i + \varepsilon_{it} \quad , i = 1,\ldots,11 \; t = 1,\ldots,6$$

Here, y is Logarithmic urbanization rate, used to measure the urban development level; x is Logarithmic dependent variables, including Per Capita GDP (yuan), the unit number of on-the-job at the end of a year(ten thousand people), social security and workers employment spending (one hundred million yuan), the consumer price index(Last year was 100); α_i is Variable intercept $i = 1,\ldots,11$; α_0 is Fixed intercept; ε_{it} is Errors, we usually assume that

random errors are independent to each other, and meet the mean zero, the same variance with assumptions.

All the dependent and independent variables are logarithmic, which is used to reduce the influence of heteroscedasticity.

E.results of model parameters estimation

We use the urbanization rate of Hubei $LOG(M)$ as Interpreted variable, the number of employees at the end of the year（ten thousand people） $LOG(G)$ and Social security and employment expenses (a hundred million yuan) $LOG(J)$ as explanatory variables. We use 11 regions as cross section unit, and the sample interval is from 2008 to 2013 to make estimation[5], the results are as follows:

TABLE III Fixed influence variable coefficient model analysis results (urbanization and the number of employees at the end of the year)

Region	Fixed intercept	Variable intercept	β Coeffici ent	Sig.
Wuhan	0.756	24.909	-3.059021	0.0000
Huangshi	-	0.611	-1.155657	0.0000
Shiyan	-	-13.352	0.210273	0.0238
Yichang	-	-6.180	0.011113	0.8511
Xiangyang	-	-8.633	0.258416	0.0003
Ezhou	-	-1.815	-0.488975	0.1646
Jingmen	-	-8.364	0.026371	0.6765
Xiaogan	-	19.498	-0.087281	0.0468
Jingzhou	-	-7.814	-0.012691	0.8721
Huanggang	-	11.795	-0.104458	0.0058
Xianning	-	-10.656	0.006263	0.9172
R-squared	0.972577			
F-statistic	74.30922			

TABLE IV Fixed influence variable coefficient model analysis results (urbanization and Social security and employment expenses)

Re gion	Fixed intercept	Variable intercept	β Coefficient	Sig.
Wuhan	0.885	2.442614	-0.393267	0.0001
Huangshi	-	1.161091	-0.341826	0.0000
Shiyan	-	-0.470071	0.084776	0.0249
Yichang	-	-0.220130	0.007139	0.8519
Xiangyang	-	-0.640887	0.232281	0.0000
Ezhou	-	-0.027368	-0.112130	0.1180
Jingmen	-	-0.457564	0.017739	0.6649
Xiaogan	-	-0.386792	-0.021608	0.5013
Jingzhou	-	-0.493181	-0.004346	0.8519
Huanggang	-	-0.482405	-0.075067	0.0036
Xianning	-	-0.425307	0.001658	0.9463
R-squared	0.973935			
F-statistic	78.29002			

IV. EMPIRICAL ANALYSIS

By the result, we will make empirical analysis for the explanatory variables $LOG(G)$ and $LOG(J)$ of fixed influence variable coefficient model analysis

A. urbanization and the number of units employed in the end of the year[6]

Seeing from the notable value of Tab3's β Coefficient, in this model, except Ezhou, Jingmen, Jingzhou, Xianning, Yichang's regression coefficients were not significantly, the regression coefficients of the other regions have all passed the 5% regression coefficients test. On the results of Analysis, the regression coefficient is negative of all regions except Xiangyang, Yichang, Shiyan, Jingmen and Xianning. The reason for this condition is that Xiangyang, Yichang, Shiyan, Jingmen and Xianning's population pressures is not so heavy compared to the rest of the regions. So the increase of population employment is still active to the promotion of urbanization. While the rest of the regions' employment population has reached saturation level and aggravating the urban pressures as population gathering area, the number of employment has hindered the process of urbanization. In this regard we can improve the urbans' comprehensive service ability and guide the surrounding area of population aggregation to alleviate the pressure of the ecological region, enhance the ability of sustainable development and stimulate the surrounding areas' urbanization process. At the same time, according to the table we can also see that the numbers of employees makes different effects in different regions. Wuhan and Huangshi have a high level of urbanization and the employment population is highly concentrated as important cities of Hubei Province. The numbers of employees perform a huge difference on their resilience of urbanization, which is mainly dependent on these areas of population distribution and industrial structure. This also shows that in the late stage of urbanization, the highly urbanization areas are very sensitive to the employment situation. As for Shiyan, Xiangfan, Ezhou and Huanggang, the employment situation doesn't make too many effects on their resilience of urbanization, they each have their respective advantages and disadvantages, the employment situation's effect has not been fully released.

B.urbanization and Social security and employment expenses[7]

Seeing from the notable value of Tab4's β Coefficient, in this model, except Yichang, Ezhou, Jingmen, Jingzhou and Xianning's regression coefficients were not significantly, the regression coefficients of the other regions have all passed the 5% regression coefficients test. On the results of Analysis, the regression coefficient is negative of all regions except Wuhan and Huangshi. The reason for this condition is that Wuhan and Huangshi's regional economic development is relatively fast, and the direction of the urbanization construction has begun to turn to the social security aspects. The increase of social security and employment expenses has positive effect on the promotion of Urbanization. Meanwhile, the other regions' development is relatively backward, the urbanization construction pays

more attention to the economic development, so the social security and the employment expense are hindering their urbanization process. In this, we suggest to consider their situation ,develop manufacturing and service industries and promote the division of labor and professional development while appropriately reduce social security and employment expenses. Then we can use the industrial structure adjustment to solve the problem, rather than simply rely on social security and employment spending. At the same time, the social security and employment expenses of different regions also have different effect for their urbanization. Although the increase of social security and employment spending has a positive effect on the urbanization of Wuhan and Huangshi, its effect is different: Social security and employment spending on the resilience of urbanization in Wuhan is obviously bigger than Huangshi. For other regions, the social security and employment spending on the resilience of the urbanization are relatively small, which is due to the backward of their economic development， their social security and employment spending is not the main contradiction. Overall, the impact of social security and employment on the urbanization of the city mainly depends on the developed level of the city. In the early time of development, we should pay more attention to economic development. Then in the later time of development, the effect of expenditure for social security and employment is increasing gradually. The better the city's economy develops, the greater social security and employment spending's depend's effect on the urbanization

V. CONCLUSION

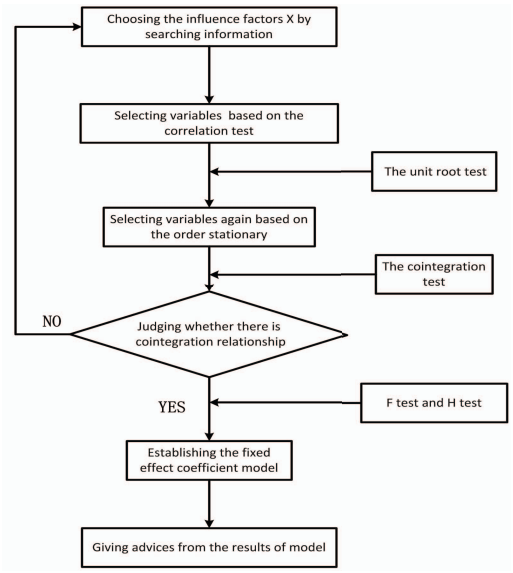


Figure I. Panel data model's flow path

Figure I is based on 12 important indicators of 11 important cities in 6years. We construct a fixed effect variable coefficient model including employment, economic development, social security and people's life by Unit root test and cointegration test. The original 12 explanatory variables are selected to retain four first order stationary variables: Per capita GDP, the number of employees at the end of the year, the social security of employment, the consumer price index of the consumer. Then we use F test and H test to select 2 variables that can establish the fixed effect coefficient model---the number of employees at the end of the year, the social security of employment, and make empirical analysis. We think that in the model of the number of employees at the end of the year and the level of urbanization, in the late stage of urbanization, the highly urbanization regions are relatively sensitive for the employment situation. In the model of urbanization level and social security and employment, the impact of social security and employment expenses on city's urbanization is mainly determined by the developed level of the city. In the early stage of urbanization, we should pay more attention on economic development and in the latter stage of urbanization, effect of expenditure for social security and employment is increasing gradually. The more the city economy developed, the greater spending on social security and employment.

VI. ACKNOWLEDGMENT

The paper is financially supported by China national natural science foundation (No.51279149, No.51479156), and Fundamental Research Funds for the Central Universities(WUT:2015IA005), self-determined and innovative research funds of WUT:2014-LX-B1-10, and Class independent innovation project of Wuhan University of Technology No2015-WL-C1-16

VII. REFERENCES

[1] Yin, Jiangbin, and Xun Li. Retrospect and Prospect of China's Population Flow and Urbanization. Urban Problems.2012, (209):23-29.

[2] Li, Wei,and Zeluo Ju.The Study on the Relationship Between Chongqing urbanization Rate and the Urban and Rural Residents' Consumption Expenditure. Journal of Chongqing Industry and Commerce University (natural science edition).2014, 31(1):16-21.

[3] Lin, Xiaojie. The Empirical Analysis of the Relationship with the Urbanization Level and Social Security of China. China Management Informationization .2014, 17(9): 87-90.

[4] Wu,Zhenxing et al. The Influence of Population, Resources and Environment for Economic Development——the Empirical Analysis Based on the Panel Data of Chinese Provinces. Mathematics Experiment and Understanding.2011, 41(12):33-38.

[5] Huang, Yong. International Comparative Analysis of the Promoting Effect from Urbanization on Economic Growth and Labor Transfer——Analysis Based on Fixed Effect Variable Coefficient Model of International Panel Data. The Financial Times. 2013,(538):128-130.

[6] Wang,Hong,and Kaichang Cui. The Empirical Study of the Relationship with China's Employment Growth and Urbanization Level. Social Sciences In Nanjing.2012, 22(8):28-48.

[7] Nagy R C, and Lockaby B G.Urbanization in the Southeastern United States: Socioeconomic Forces and Ecological Responses Along an Urban-rural Gradient.Urban Ecosyst.2011,(14):71-86.

A risk probability model of study large vessel navigation with wind and water flow

Li Zhenping
Department of Mathematics Luoyang Institute of Science and Technology,Luoyang , China
lizhp2000@126.com

Zhang Shesheng, Gui Yufeng
Mathematical model of the association
Wuhan University of Technology
Wuhan, China guiyufeng@hotmail.com

***Abstract*—With the development of shipping industry, the large vessel safety of sailing in channel becomes one of the most important issues of concern. This paper considers the influences of drift by wind and current, ship's drift distance combines normal distribution, analyzing the probability distribution of ship, and predicting the risk of ship navigation. It provides some reference for the ship's officer to judge the vessel movement and guarantee the safety of navigation..)**

Keywords-component; navigation safety; ship grounding; drift probability density distribution

I. INTRODUCTION

Ports have the important position and effect in national economy and opening to the outside, the allocation plan of ports continuously extend nowadays, however, it is common that there are many vessels accidents occur in channel from these years' data statistics. Due to the navigable waters are limited, and numerous vessels, the lower speed reduce the maneuver-ability of the ship, with the influences of wind, current and waves, leading to ship safety accidents. From many statistical information of these years, ships collision and grounding accidents have a great relation with wind and current [1],because ships inwards and outward the port fits national< <Design Code of General Layout of Sea Port> >[2] by using many research production of correlative technique literature, regarding the probability of ships navigate as normal distribution when calculate the deviation probability of ship in channel, the standard deviation is the length of the ship, the center of the channel is the mid-value position of the normal distribution, as shown in Figure 1..

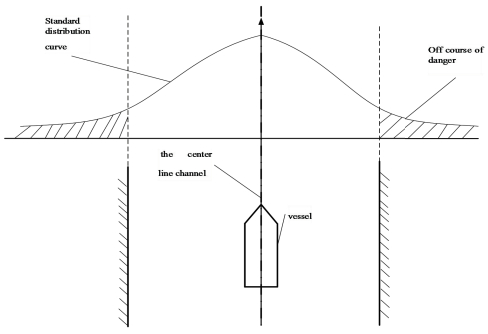


Figure 1. ship navigation risk probability density distribution.

II. THE ESTABLISHMENT OF DRIFT DISTANCE MATHEMATICAL MODEL

Ships navigate in the channel will follow the direction of rudder when there are no wind and current, they will produce drift distance caused by wind and current[3] ,meanwhile having the effect on the width of track, and then making the change of navigation width. From every kind calculation of track and drift distance influenced by wind and current, we gain the probability density of vessel navigation risks after the influence of wind and water flow.

.The angle between fore and aft line and channel center line is yaw angle α ,the angle between flow direction and channel center line is the angle of current β . The drift distance (Δ β 1) of any ships influenced by all kinds of currents equation as follow[4] :

$$\Delta B_1 = S \cdot \frac{V\sin\alpha + U\sin\beta}{|V\cos\alpha + U\cos\beta|} \tag{1}$$

There into: S—Calculation of the channel length (m);

V—Ship speed (m/s);

U—Flow rate (m/s).

The ship navigation drift distance (Δ β 2) effected by wind equation as follow [5]:

$$\Delta B_2 = K \cdot \sqrt{\frac{B_a}{B_w}} \cdot e^{-0.14V_S} \cdot V_a \cdot S \cdot \frac{\sin\alpha_f}{|V\cos\alpha + U\cos\beta|} \tag{2}$$

There into：

$$K = \sqrt{\frac{\rho_a \cdot C_a}{\rho_w \cdot C_w}} \tag{3}$$

The coefficient of general value 0.038~0.041；

B_a Upper hull waterline wind age area（m2）；

B_w Hull wind age area under water line(m2)，；

V_S Ship speed in wind（m/s）；

V_a Relative speed of wind（m/s）；

α_f The angle between true wind direction and the center line of the channel.

Total drift

$$\Delta B = \frac{SV\sin\alpha + SU\sin\beta}{|V\cos\alpha + U\cos\beta|} + K\sqrt{\frac{B_a}{B_w}}\frac{e^{-0.14V_S}V_a S\sin\alpha_f}{|V\cos\alpha + U\cos\beta|} \tag{4}$$

III. THE PROBABILITY DENSITY DISTRIBUTION UNDER THE INFLUENCE OF THE WIND

After the impact of wind and current, the risk probability density distribution of no wind and current will be corrected

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.120

by wind and current, as shown in Figure 2, the mid-value of normal distribution isn ’ t in the center line, the distance difference is the wind and current drift,the size is the drift caused by wind and current, the function is:.

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

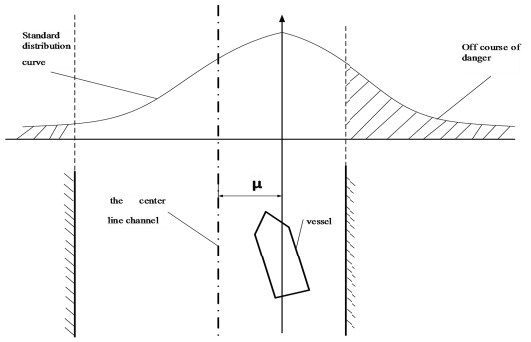


Figure 2. Ship navigation risk probability density distribution after correction.

IV. DRIFT MATHEMATICAL MODEL ARITHMETIC

According to the establishment of drift mathematical model, using the calculation method based on numerical calculation method based on numerical calculation software MATLAB [6] , calculate the risk probability of ship due to the length of the ship and drift. According to the risk probability differential get corresponding influence factors sensitivity of probability of risk..

According to the established mathematical navigate in the channel

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

Here μ is the total amount of drift , σ is the ship length, said the ship in the channel obey Gaussian distribution which has the parameters μ and σ .

Function f(s) is symmetrical about μ , f(s) is maximal when μ ,means no wind and current, the probability of ship near the center line of the channel is the largest. When μ fixed, the smaller of σ ,the moregraph pointed; on the contrary, the bigger of σ , the graph is flat. When σ fixed, μ changed, mid-channel corresponding is no longer maximum, the channel value of is changed with different position, this reflects the “drift μ” and the “length σ” of the ship have an influence on ships navigation probability density..

For the risk probability of ship P

$$P=1-P\{-x_0<x<x_0\}=1-\int_{-x_0}^{x_0}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \tag{7}$$

Hre x_0 is the half width of the channel also the limit drift.

Ships in channel can be expressed as the probability density function f(x) integral in [-x0,x0] ,the remaining is the risk probability, apparently, the more wide of the channel, the lower risk.

Method to calculate integral expression.

$$S=\int_{-x_0}^{x_0}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \tag{8}$$

Taking n=1000 points integral express in [-x0,x0], get p_j, among them density of the ship

$$p_i=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

According to the value Pi of the requested, for , from 1 to 1000 sum of the ship safety probability S,

$$S=h\sum_{i=1}^{n}p_i=h\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Where is the step length.

Difference method is used to calculate integral expression directly changes the approximate numerical differential problem concept is straight forward, due to the choice of n=1000, the calculation accuracy meets the needs of the study.

$$P=1-P\{-x_0<x<x_0\}=1-\int_{-x_0}^{x_0}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \tag{9}$$

According to different influencing factors, in the independent variable interval by interval H even take. The sensitivity of risk probability is T=P’, in this paper, with the following discrete formula.

$$T=\frac{P(i+1)-P(i)}{H} \tag{10}$$

V. EXAMPLE ANALYSIS

According to the port relative information, choosing the width of the channel, discussing the ships length , total amount of drift on the influence of the ship in channel navigation risk probability, then calculated ship risk probability distribution under different wind and ship length influences. In order to find further illustrate the trend of change, two kinds of impact sensitivity are analyzed respectively and find out the change trend of impact..

According to the port related information, channel width x is150m/1000m, the total amount of drift is 0, the length σ of the variable range is100m/1000m—300m/1000m, taking interval H=5m/1000m.

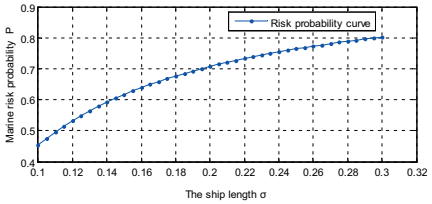


Figure 3. length's influence on the risk probability of the ship

MATLAB numerical method of calculating the ship length realized for different risk probability calculating, as you can see from the chart, along with the rising of the ship ' s length σ , ship ' s risk probability P also corresponding increasing (as show in figure 3). Graphics the shape of the curve gradually flatten out, when the ship length growth continued, risk probability of slow growth, the curve is convex.

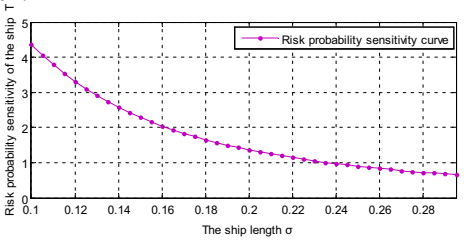


Figure 4. The influence of ship length on risk probability sensitivity.

From the change of the ship length σ of the risk probability can be seen on the influence of sensitivity, probability sensitivity decreases (as shown in figure 4),so the ship in small size, the difference on the ship ' s length impact on the risk probability of the navigation in the channel is more obvious, when the ship ' s length is long, the influence of ship risk probability is not significant.

According to the port relevant information, the channel width x is 150m/1000m, ship lengthis200m/1000m, the total amount of drift variable range of 0 — 50m/1000m, taking interval H=1m/1000m.

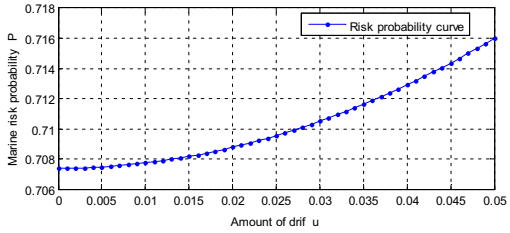


Figure 5. scatter is the impact on the risk probability

According to the relevant data, fixed the ship length and channel length, using numerical

calculation, the total amount of drift on ship risk probability P associated with large amounts of data, from the chart, along with the increase of drift, marine risk probability corresponding increase (as shown in figure 5).the shape of the curve is steep gradually, when the total amount of drift continue increasing, risk probability change steep growth, curve is concave.

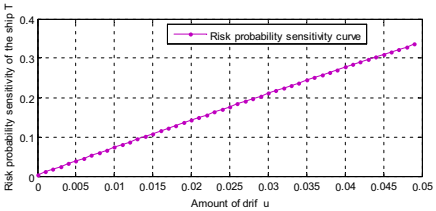


Figure 6 . Scatter is the impact on the risk probability sensitivity

Total amount of drift on marine risk probability P can be seen on the influence of sensitivity, with the increase of total amount of drift, the sensitivity is bigger and bigger, it shows that the possibility of ship ' s dangerous rate along with the increase of ship drift, it is disadvantageous to navigation. Ship as close to the shore, the greater the risk probability changes faster, the risk increase faster, (as shown in figure 6). As a result, the total amount of drift impact on marine risk probability is extremely significant.

VI. RESULT ANALYSES

Analyzing the factors of drift through the model of calculation and analysis, we can see that drift effects on the risk probability of ship is very significant, when the ship navigating under the influence of wind and current, and the drift is small, can proper fine-tuning or don ' t do adjustment when drift volume continues to increase, the officer on duty should pay attention to, according to the actual situation, use the corresponding rudder angle, to ensure safety of navigation. Due to the low speed of ship entering the port, results in the decrease of steering, it difficult to control the ship if the wind drift amount is larger, this situation is threatening the safety of the ship, especially the ship grounding, according to the actual condition the officer can use the method of extra help rudder to ensure the ship safety.

Using the corresponding data obtained from the method of numerical calculation, also as the growth of the length of the ship, marine risk probability P is bigger. Compared with the amount of drift factors, ship length of risk probability impact is slower. For smaller length difference ship risk probability difference is not big, but as for a greater difference of ship length, the difference of risk probability is very big. So before sailing, we must have a basic understanding of the ship characteristics. As for the ship staff should be accurately grass the ship dangers of driving in the waterway, knowledge of ship maneuvering effect and the effectiveness of the drift correction

VII. CONCLUSIONS

Large vessels inward and outward port, in addition to the problem of speed slower, steering is reduced and their own size, etc. Also have the environment problems of numerous ships, limited of the channel, wind and current, these factors is very important to the safety of the ship. This paper uses the method of numerical calculation. Emphatically analyzed " ship length " and " wind and current " impact on ship safety. Analysis results provide the reference for driver to protect the safety of ship navigation moment, and has certain theoretical value in reducing ship grounding, meanwhile, it can also be used for reference for channel planning. This paper involves the channel width but none of the factors as the ship ' s draft, at the same time, the angle of the ship certainly shorten the navigable areas, this problem will require further research reference.

ACKNOWLEDGMENT

The paper is financially supported by China national natural science foundation (No.51139005), Humanity and Social Science foundation of Ministry of Education of China(Grant No.12YJAZH022) and Foundation of Department of Commerce of Hubei, China(Grant No.HBSW-2014-01).

REFERENCES

[1] [1]Li Dongyang, Zhu Huaiwei, Zheng Zhongyi. The ship ran aground cause analysis. Traffic Word, 2009(11), pp. 276-279.

[2] [2]Miao Hui, Sun Yingguang. Discuss on part of clauses of Harbor Designing codes of General Layout JTJ211-99, 2006(10), pp. 329-334.

[3] [3]Liu Mingjun, Li Bing. Study on probability density distribution of ship impact against bridge after the influence of wind and water flow, China Science Paper on Line. 2010(4), pp. 1-4.

[4] [4]Gu Yi. Channel width algorithm based on scatter is analyzed. China Water Transport (Theory). 2006(11), pp. 23-24.

[5] [5]Liu Mingjun, Lv Xidao. Ship required navigation wide curve modeling. Wuhan University of Technology Journal (Traffic Science and Engineering). 2006(1), pp. 178-179.

[6] [6]Zou Chunming. Ship safety navigation based on MATLAB7.0 width calculation. China Water Transport. 2008 (5), pp. 9-10..

A research ship characteristic length model based on statistical theory

Xuefei Zhang, Xin Chen
Wuhan University of Technology, Wuhan, 430070,China
e-mail: ChenXin@qq.com

Zhenli Sun, Shesheng Zhang Yuguang Li
Wuhan University of Technology, Wuhan, 430070
e-mail: Liyuguang@qq.com

***Abstract*— Ship characteristic length directly affects the development plans of the company. Based on the statistical theory, a data categorical statistical model is build for ship length research. The sub division calculation method is given, and the statistical subsection regression is calculated between ship length, time and GDP. At last, The data of ship characteristic length is forecasted for 2020 year.**

Keywords-ship, characteristic length, statistics, stochastic distribution

I. INTRODUCTION

The design of ship manufacturing platform must consider the length of the ship. For example, 500m long ship is not manufacturing on the 10m long boat platform, and 10m long boat manufacturing will cause high cost on the 500m long ship platform [1]. Method of management of boat is not the same one as huge ship. The loading method and the boat ship docked, ship transport routes, the goods have different [2]. So the study on the ship characteristic length has strong actual background of [3].

Ship classification and recognition have great significance for the monitoring and management of marine transportation, as well as an important part of the SAR ocean application. Based on the structural characteristics of the commercial ship, paper [4] presents a commercial ship classification algorithm. Through the synchronous experiment in the East China Sea test area, the average COSMO-SkyMed image classification commercial ship classification algorithm achieves an accuracy of 89.94%. Also ship classification can be done by using ship characteristic length. But it can't work in practice. But in practice, due to river width cannot become larger as time become larger, thus, the ship characteristic length cannot become large enough in the rivers [5]. In the sea, the ship characteristic length may become larger. This is means that the research of ship characteristic length must consider different length of ship. [6].

This paper will study ship characteristic length, and establish a statistical analysis model, find out the characteristic length changed with time..

II. BASIC ANALYSIS

In this paper, data is collected from the international shipping network, Baidu or other open website. Each data contains the captain, the beam, manufacturing time, and the corresponding economic data, such as GDP etc. Then data arranged according to the following methods

The ship characteristic length is arranged from small to big, and then divided into P parts. K=1 is subscript for small characteristic length part, and k=P is subscript for largest one

Secondly, data is arranged according to the year and captain from big to small order. In the same way, data is divided into P parts on a yearly basis. K=1 is subscript for small characteristic length part, and k=P is subscript for largest one.

Fig.1 shows than ship length varied with time from 2000 to 2012's. From figure, we find, the number of collected data is more after 2006. That means the number of big ship is different for different year. It will result in the statistical error, especially the time error in regression equation. We will use data divided method to reduce above statistical error.

The ship characteristic length is divided P parts. Let yk represented kth part of ship characteristic length (unit is m). $x_1=t$ Is time(unit is year), x_j,j=2,3,4,..,M are economic statistical varied. Such as GDP, average wage, et al. Let

$$y_k = a_{k0} + a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kM}x_M + \varepsilon_k \quad (1)$$

.k=1,2,3,..,P

Here is normal error distribution with mean zero. By using data (ykj,x1s,...,xMs), s=1,2,...,Nk. We have:

$$y_{ks} = a_{k0} + a_{k1}x_{1s} + a_{k2}x_{2s} + \ldots + a_{kM}x_{Ms} + \varepsilon_{ks} \quad (2)$$

.k=1,2,3,..,P, s=1,2,,,.Nk

Let

$$X_k = \begin{pmatrix} 1 & x_{11} & \ldots & x_{M\mathrm{Nk}} \\ 1 & x_{12} & \ldots & x_{M\mathrm{Nk}} \\ & & \ldots & \\ 1 & x_{1\mathrm{Nk}} & \ldots & x_{M\mathrm{Nk}} \end{pmatrix} \qquad A_k = \begin{pmatrix} a_{k0} \\ a_{k1} \\ \\ a_{kM} \end{pmatrix} \quad (3)$$

$$Y_k = \begin{pmatrix} y_{k1} \\ y_{k2} \\ \\ y_{kNk} \end{pmatrix} \qquad \Omega_k = \begin{pmatrix} \varepsilon_{k1} \\ \varepsilon_{k2} \\ \\ \varepsilon_{kNk} \end{pmatrix}$$

We have matrix formula:

$$Y_k = X_k A_k + \Omega_k \quad (4)$$

It has solution of minimum error

$$A_k = (X_k^{\,T} X_k)^{-1} X_k^{\,T} Y_k \quad (5)$$

The sum of absolute error are

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.121

$$|| \Omega_k ||=| \varepsilon_{k1} | +...+ | \varepsilon_{kNk} | \quad (6)$$

Here k=1,2,..,P. The ship characteristic length is divided P parts, we have correlation matrix between P parts.

$$R = \begin{pmatrix} R_{11} & R_{12} & ... & R_{1P} \\ R_{21} & R_{22} & ... & R_{2P} \\ ... & ... & ... & ... \\ R_{P1} & R_{P2} & ... & R_{PP} \end{pmatrix} \quad (7)$$

where

$$R_{ij} = \frac{E(Y_i Y_j) - E(Y_i)E(Y_j)}{\sqrt{D(Y_i)D(Y_j)}} \quad (8)$$

III. TCALCULATION RESULTS

.We know t is time, x=GDP of China, the regression equation is

$$L = 32467 - 16.13t + 0.00073393x \quad (9)$$

From this formula， the time coefficient is (-16.13<0). That means the ship length will become shorter with time. But in practice, the ship length is increase with time. The reason is a lot of small boat after 2006. So that dividing length data may be improving result. By using this ideal, the data is divvied 5 parts. According to the random theory, the ship length is Yj random varies, its average value of Yj part is shown in the table I. The variance is also shown in the same table. From table, we can find that the average is increase from 91.968 to 293.46, and variance is decease from 3016.4 to 396.87. Fig.2 shows the variance varied with average. From figure, the variance is decease with average, and the decease speed is increase as average increase.

Table I Average E(Yj) and varance D(Yj)

j	1	2	3	4	5
$E(Y_j)$	91.968	142.34	192.72	243.09	293.46
$D(Y_j)$	3016.4	1724.5	857.32	414.83	396.87

The correlation coefficient matrix is shown in the Table II, From tale, we find, R12=0.99278, R45=0.73863, the correlation coefficient of small boat is larger than the one of big ship.

Table II correlation coefficient matrix

	1	2	3	4	5
1	1	0.99278	0.94038	0.6796	0.0074543
2	0.99278	1	0.97438	0.76268	0.12733
3	0.94038	0.97438	1	0.8886	0.34713
4	0.6796	0.76268	0.8886	1	0.73863
5	0.0074543	0.12733	0.34713	0.73863	1

We can calculate matrix X and vector Y from given data, and the regression equations are easy obtained from above data:

$$\begin{aligned} L_1 &= 10365 - 5.1293t - 0.00025171x \\ L_2 &= -3223.4 + 1.6918t - 0.0019946x \\ L_3 &= -19406 + 9.812t - 0.0048579x \\ L_4 &= -4930.5 + 2.5867t - 0.00095029x \\ L_5 &= -4874.9 + 2.5809t - 0.00029937x \end{aligned} \quad (10)$$

Here Lk is ship length(m),k=1,2,3,4,5, t is time(year),x is China GDP(GRMB). L1 is represented small ship length, and L5 is represented large ship length. From above formula, all coefficient of x is less than zero, and the values are very small. That means GDP is weak affect ship length. Next discuss time coefficient. From above formula, except L1, all coefficient of time t is larger than zero. That means the ship length is increase with time. Fig.3 shows the time coefficient varied with k. From the figure, the time coefficient is largest when k=3. Fig.4 shows ship’s length varied with time(year), in the figure, (a) small boat length varied with time; (b) big ship length varied with time.

The regression error for different ship length is shown in the table 3. In the table, Y is ship’s length measured. L is ship’s length calculated from regression formula. k is represented kth part of ship’s length. Sum |L-Y|k is total error for part k. Average |Lj-Yj| is average error for part k. Lk is average ship length for part k. Relative |Lj-Yj|/Lk is relative error for part k. From the table, we find, the regression error is largest for small boat, and smallest for big ship. The relative error is only 6.148%.

Table 3 regression error for different ship length

k	Sum	Average	Relative
1	1341.9	33.547	0.36477
2	1073.6	26.84	0.18856
3	775.46	19.387	0.1006
4	699.57	17.489	0.071946
5	721.69	18.042	0.06148
sum	4612.2	23.061	0.15747

Table 4 predict the average ship largest length varied with year between 2012 to 2021. From table, we find that the lengths will be increase to 472.49 in 2020.

Table 4 ship length varied with time

year	2012	2013	2014	2015	2016
L	346.65	360.63	374.61	388.59	402.58
year	2017	2018	2019	2020	2021
L	16.56	430.54	444.5	458.51	472.49

IV. CONCLUSIONS

In the above discuss, we built a statistical model of the relation between ship length and shipbuilding, and also GDP of China. The small boat data is to much to make large error. When the data is partitioned into several classes, we found that this way can improve the regression effect, and can reflect the ship length increase with year. The data is divided five parts. From the calculation results, we found, the small boat length is not almost varied with time, but big ship's characteristic length is positively correlated with years. Statistics show that, the characteristic length of different ship classification with years of change is different. From above discuss, we find how much part divided will be need us consider in the research.

ACKNOWLEDGMENT

The paper is financially supported by China national natural science foundation (No.51139005)

REFERENCES

[1] .Kai Sun, Research on large container quay operation process on the ship's[J], Maritime China, 2015(01),60-62.

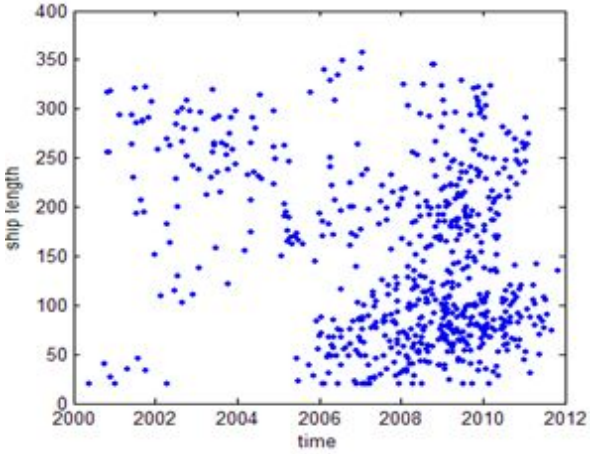
[2] [2]Xiao Li, Analysis of the effect of 400000 ton ship to the port of importing iron ore usiness, China waterway, 2015(01),52-53.

[3] [3] LIU Zhao , LIU Jing-xian , ZHOU Feng, Behavior characteristics of vessel traffic flow and its realization in marine traffic organization[J], Journal of Dalian Maritime University, 2014(02),22-26.

[4] [4] Jiang Sheaofeng, et al., algorithm for merohant ship images based on structural feature analysis[J], Remote Sensing Technology and Application, 2014,29(4),607-615.

[5] [5] Xin Chen, Shengping Jin, Shesheng Zhang, Dan Li, A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil [J], Journal of Algorithms & Computational Technology,2015,Vol. 9 No. 2, 163-174.

[6] [6] Xin Chen, Mengyu Li, and Shesheng Zhang , A Parallel Algorithm of Non-Linear Fluid-Solid Coupling Problem for Hydrofoil [J], Journal of Algorithms & Computational Technology,2014,Vol. 8 No. 3, 249-266..



Ship length varied with time(year).

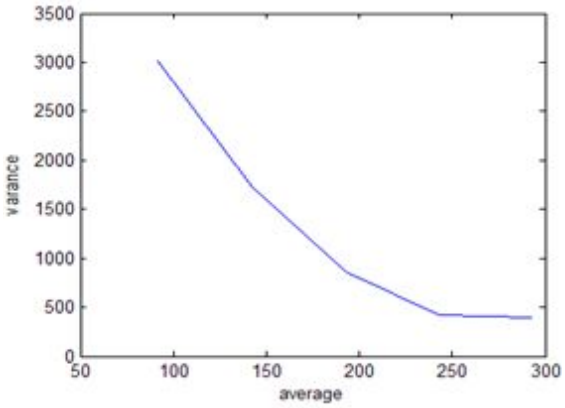


Figure 1. Variance varied with average

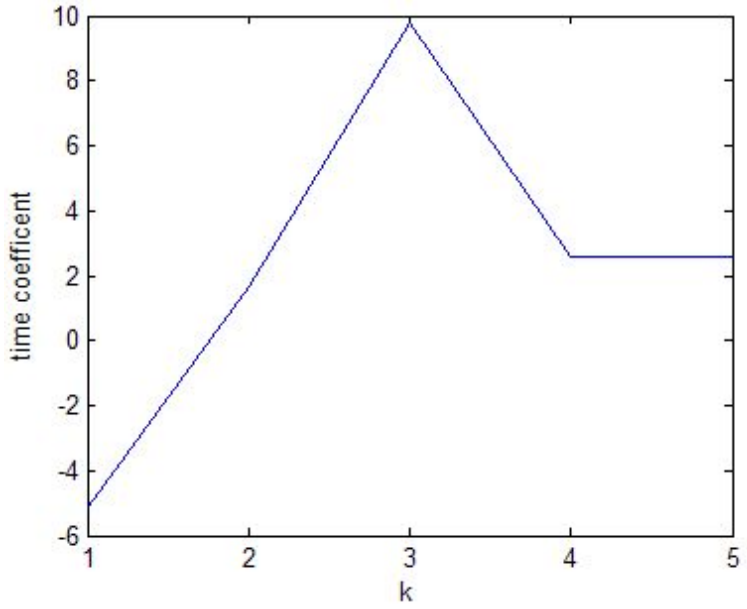


Figure 2. time coefficient varied with k

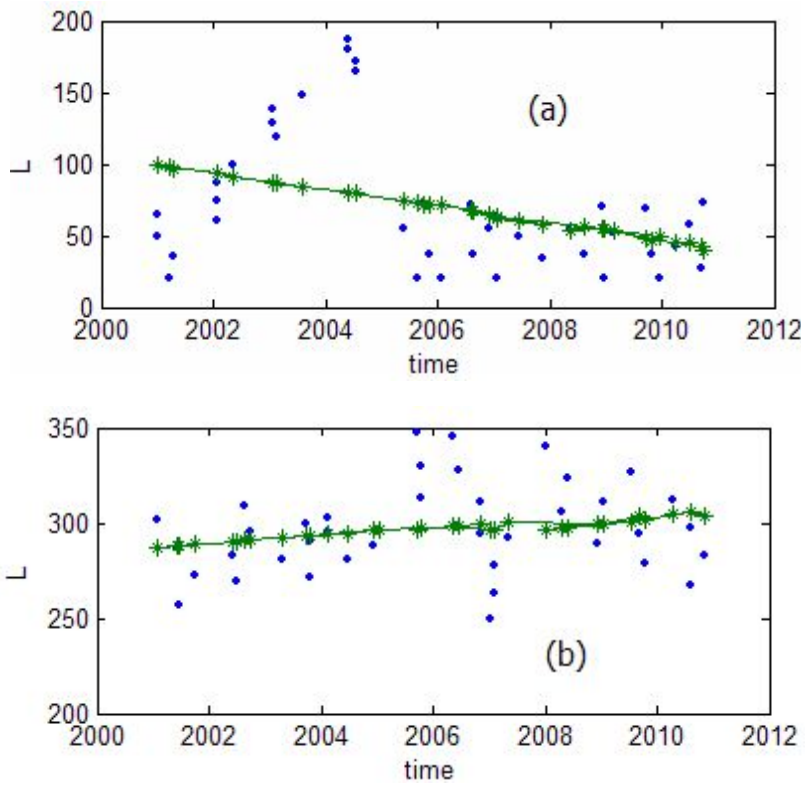


Figure 3. Ship’s length varied with time(year),(a) small boat;(b) big ship

Chaotic oscillation suppression of the interconnected power system based on the adaptive back-stepping sliding mode controller

HUANG Wen-di, MIN Fu-hong,WANG Zhu-lin, CHU Zhou-jian
School of Electrical and Automatic Engineering
Nanjing Normal University
Nanjing, Jiangsu, China
e-mail: minfuhong@njnu.edu.cn

***Abstract*—The second-order interconnected power system is investigated in this paper, and the dynamic characteristics of the system under the disturbance of load and electromagnetic power are analyzed through the Lyapunov exponent and Poincaré maps. A adaptive back-stepping sliding mode controller is designed to control the chaotic motion in the power system. Theoretical analysis and numerical simulation results show that the controller can effectively suppress the chaotic oscillation.**

Keywords-chaotic oscillation; interconnected Power System; back-stepping sliding mode control

I. INTRODUCTION

The security and reliability of the electric power system are vital for social civilization and national economy. However, with the technology of ultra-high voltage, high-capacity and long-distance transmission lines growing fast in the modern power grid, the difficulty of maintaining power system dynamic stability gradually increased. Chaotic phenomenon is inevitable in power systems, and several power failures have been caused due to chaotic oscillation, which have aroused the great attention of the public. To control a relatively new thing, people need to constantly seek all kinds of new methods of analysis and control[1-3].

Recently, the research of chaos in power system has become a hot topic. For example, the difference between chaotic oscillation and other power system instabilities was studied [4-5]. In[6], the chaotic characteristic in two-machine interconnected power system was carried out and found that the chaotic oscillation would emerged in the system when the amplitude of periodical load disturbance increased gradually. At the basis of [6], the disturbance of electromagnetic power is considered as another parameter, which described the route to chaos with the amplitude of electromagnetic power disturbance increased [7]. Different chaotic oscillations might cause a huge impact to the stability of power system, and then the chaotic oscillation should be controlled to avoid angle instability and voltage collapse.

Nowadays, there have been a lot of chaos control methods, and each has own advantages and disadvantages and different application conditions.The sliding mode control (SMC) is becoming more and more popular, which is valued for the robust accommodation of uncertainties and the ability to reject disturbances. There is a gap between the sliding mode control and the application in real world. The gap is chattering, which exists in practice. How to eliminate the chattering phenomenon while keeping the robustness property of SMC is an important problem to be resolved for the wide application. The fuzzy fast terminal sliding mode controller was proposed to stabilize the power system to synchronization status based on equivalent control [8]. A sliding mode controller was described based on the relay characteristic function [9] and simulation results clarified that the controller can shorten the control time and reduce the possibility of the impulse response.

The main theme of this paper is investigated the chaotic phenomenon in interconnected power system under the disturbance of load and electromagnetic power. The ranges of disturbance amplitude and frequency for the chaotic oscillation are studied in the system. Moreover, the adaptive back-stepping sliding mode controller is designed to make the system quickly and smoothly reach the expected target.

II. SYSTEM MODEL AND DYNAMIC ANALYSIS

A. System Description

A dynamic model of simple 2-order interconnected power system model is considered, which is widely known and used as a benchmark example in the literature. The schematic diagram is depicted in Fig.1.In this diagram,1 is equivalent machine of system1. 2 is equivalent machine of system2. 3 is main equivalent transformer of system1. 4 is main equivalent transformer of system2. 5 is load. 6 is breaker. 7 is the linking lines between two systems. The dynamics of the system can be expressed by the following nonlinear differential equations[9]:

$$\begin{cases} \dfrac{d\delta(t)}{dt} = \omega(t), \\ \dfrac{d\omega(t)}{dt} = -\dfrac{1}{H}[P_s \sin(\delta(t)) + D\omega(t) - P_m + P_k \cos(\alpha t)\sin(\delta(t) - P_e \cos(\beta t)], \end{cases} \tag{1}$$

where $\delta(t)$ is the angle and $\omega(t)$ is speed of the generator

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.122

rotor. H is the inertia constant. D is equivalent damping coefficient. P_m is the mechanical power. P_s is the electromagnetic power. P_k, P_e are the amplitude of electromagnetic power disturbance and load disturbance. α, β are the frequency of electromagnetic power disturbance and load disturbance , respectively.

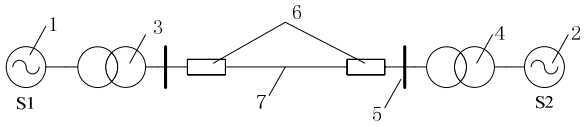


Figure1. Simple interconnected power system.

The dynamics equation(1) from [9] can be rewritten as:

$$\begin{cases} \dfrac{dx_1}{d\tau} = x_2, \\ \dfrac{dx_2}{d\tau} = -\sin x_1 - \lambda x_2 + \rho + \mu\cos(\gamma\tau) - \sigma\cos(\eta\tau)\sin x_1, \end{cases} \tag{2}$$

where $\lambda = D/\sqrt{HP_s}$, $\rho = P_m/P_s, \sigma = P_k/P_s$, $\mu = P_e/P_s$, $\eta = \alpha\sqrt{H/P_s}, \gamma = \beta\sqrt{H/P_s}, \tau = t\sqrt{P_s/H}$.

In practical engineering applications, the system load is allocated to each generator on average, and all the coefficients are given as: $\lambda = 0.4, \rho = 0.2, \mu = 0.02$.

Then, plugging the coefficients into formula(2), new formula can be obtained:

$$\begin{cases} \dfrac{dx_1}{d\tau} = x_2, \\ \dfrac{dx_2}{d\tau} = -\sin x_1 - 0.4x_2 + 0.2 + 0.02\cos(\gamma\tau) - \delta\cos(\eta\tau)\sin x_1. \end{cases} \tag{3}$$

B. Chaotic Oscillation Analysis

The previous studies mainly based on the condition that two frequencies of the disturbance are equal in [9]. In fact, there have abundant nonlinear dynamics characteristics when the frequencies are different. Therefore, δ is set as 1.3 in this paper and the whole process of system is observe d by changing the k in formula(4).

$$\begin{cases} \dfrac{dx_1}{d\tau} = x_2, \\ \dfrac{dx_2}{d\tau} = -\sin x_1 - 0.4x_2 + 0.2 + 0.02\cos(0.8\tau) - 1.3\cos(0.8\cdot k\tau)\sin x_1. \end{cases} \tag{4}$$

The Lyapunov exponent spectrum with k varying is shown in Fig.2. Obviously, when k moves into the shaded parts, chaotic oscillation occurs. The phase maps with different k are drawn as Fig.3 and the corresponding Poincare sections as Fig.4. Periodic, quasi-periodic, chaos of rich and complex nonlinear behaviors of the system is presented in Fig.3 and Fig.4. From Fig.3 and Fig.4, chaotic oscillation exists in the interconnected power system when k is 2. 4269. It is necessary to design a suitable controller to remove the hidden trouble.

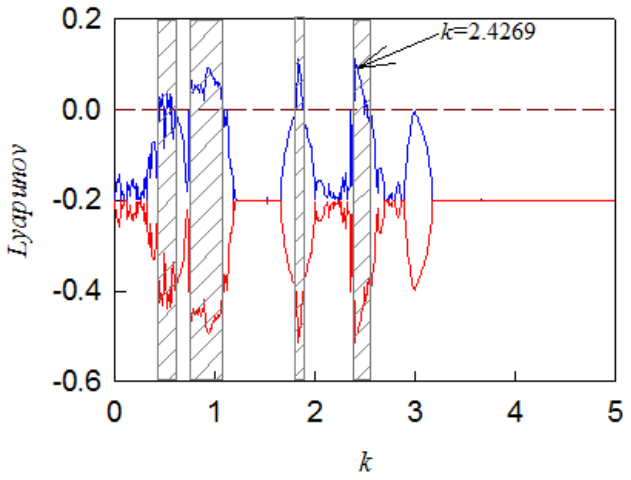


Figure 2. Lyapunov exponent with the change of k.

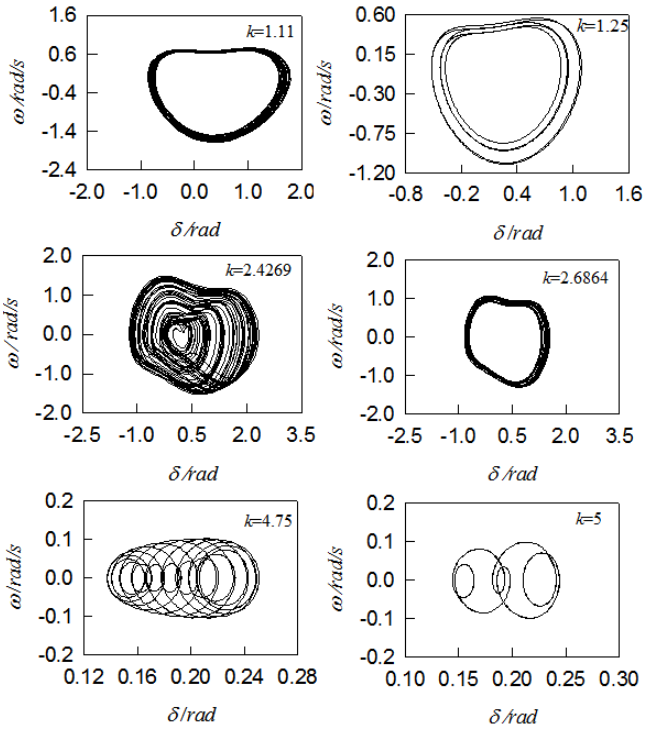


Figure 3. The phase maps of attractor

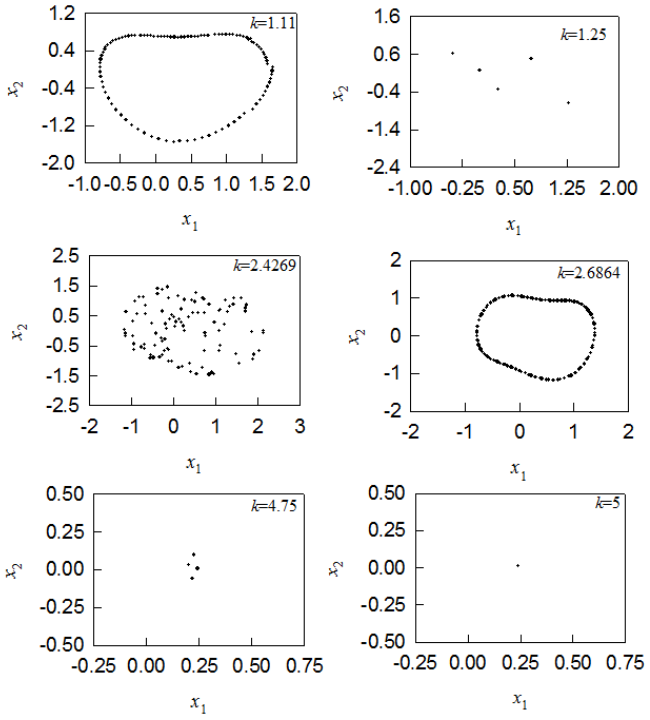


Figure 4. The Poincaré maps of attractor.

III. ADAPTIVE BACK-STEPPING SLIDING MODE CONTROL

In this paper, the nonlinear controller is devised by way of the adaptive back-stepping sliding mode control method, which not only retains the advantages of sliding mode control method, also has own feature. The basic idea of this method is divide the original complex nonlinear system into many subsystems and the order of each subsystem is below the order of original system. Lyapunov function and intermediate virtual control for each subsystem are designed to finish all the projects of control law. Especially, this method can be widely applied to the system which have total uncertainties, because the method can estimate the upper bound of the total uncertainties accurately. The detailed processes of the design are given as follows[10]:

A. System Description

Let's assume the controlled object can be obtained:

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = Ax_2 + Bu + F, \end{cases} \tag{5}$$

while F denotes the total uncertainties,

$$F = \Delta A x_2 + \Delta B u + d(t), \tag{6}$$

where ΔA and ΔB are parameters uncertainties of the total uncertainties and $d(t)$ is external disturbance. Due to the changes of uncertain part and external disturbance are very slow, so we think that:

$$\dot{F} = 0. \tag{7}$$

In practice, what is really difficult to choose a positive constant M and make $|F| \le M$. Therefore, adaptive theory should be adopted in design of the controller to estimate F.

B. Design of the Controller

Step 1:

First of all, given a position instruction x_d, and let tracking error $z_1 = x_1 - x_d$. Definite the Lyapunov function:

$$V_1 = \frac{1}{2} z_1^2, \tag{8}$$

setting z_2 as virtual control, the corresponding expressions are $z_2 = x_2 + c_1 z_1 - \dot{x}_d$, $\dot{z}_1 = x_2 - \dot{x}_d = z_2 - c_1 z_1$.

The derivative of V_1 along with equation (8) is

$$\dot{V}_1 = -c_1 z_1^2 + z_1 z_2. \tag{9}$$

Defining the switching function:

$$\sigma = k_1 z_1 + z_2, \quad (k_1 > 0). \tag{10}$$

Because of $\dot{z}_1 = z_2 - c_1 z_1$, so $\sigma = k_1 z_1 + z_2 = k_1 z_1 + \dot{z}_1 + c_1 z_1$ $= (k_1 + c_1) z_1 + \dot{z}_1$ and $k_1 + c_1 > 0$. Obviously, if $\sigma = 0$, thus $z_1 = 0, z_2 = 0$, $\dot{V} \le 0$. Turn to the next step.

Step 2:

Base on the adaptive theory to select the Lyapunov function:

$$V_2 = V_1 + \frac{1}{2}\sigma^2 + \frac{1}{2\gamma}\tilde{F}^2, \tag{11}$$

where $\hat{F}$ is the estimate value of F, so error of estimation is $\tilde{F} = F - \hat{F}$, γ is a positive constant.

The derivative of V_2 along with equation (11) is

$$\begin{aligned} \dot{V}_2 &= \dot{V}_1 + \sigma\dot{\sigma} = z_1 z_2 - c_1 z_1^2 + \sigma\dot{\sigma} - \frac{1}{\gamma}\tilde{F}\dot{\tilde{F}} \\ &= z_1 z_2 - c_1 z_1^2 + \sigma(k_1 \dot{z}_1 + \dot{z}_2) - \frac{1}{\gamma}\tilde{F}\dot{\tilde{F}} \\ &= z_1 z_2 - c_1 z_1^2 + \sigma(k_1(z_2 - c_1 z_1) + A(z_2 + \dot{x}_d - c_1 z_1) + Bu + F - \ddot{x}_d + c_1 \dot{z}_1) - \frac{1}{\gamma}\tilde{F}\dot{\tilde{F}} \\ &= z_1 z_2 - c_1 z_1^2 + \sigma(k_1(z_2 - c_1 z_1) + A(z_2 + \dot{x}_d - c_1 z_1) + Bu + \hat{F} - \ddot{x}_d + c_1 \dot{z}_1) - \frac{1}{\gamma}\tilde{F}(\dot{\hat{F}} + \gamma\sigma). \end{aligned} \tag{12}$$

Devising the controller:

$$\begin{aligned} u = B^{-1}(&-k_1(z_2 - c_1 z_1) - A(z_2 + \dot{x}_d - c_1 z_1) - \hat{F} \\ &+ \ddot{x}_d - c_1 \dot{z}_1 - h(\sigma + n\,\mathrm{sgn}(\sigma))), \end{aligned} \tag{13}$$

where h and n are positive constants.

Choosing the adaptive law:

$$\dot{\hat{F}} = -\gamma\sigma, \tag{14}$$

substituting equation (13) and (14) into equation (12), and the following equation is given

$$\dot{V}_2 = z_1 z_2 - c_1 z_1^2 - h\sigma^2 - hn|\sigma|. \tag{15}$$

Select

$$Q = \begin{bmatrix} c_1 + hk_1^2 & hk_1 - \frac{1}{2} \\ hk_1 - \frac{1}{2} & h \end{bmatrix} \tag{16}$$

where $z^T = [z_1 \; z_2]$, therefore

$$z^T Qz = [z_1\ z_2]\begin{bmatrix} c_1 + hk_1^2 & hk_1 - \frac{1}{2} \\ hk_1 - \frac{1}{2} & h \end{bmatrix}[z_1\ z_2]^T$$
$$= c_1 z_1^2 - z_1 z_2 + hk_1^2 z_1^2 + 2hk_1 z_1 z_2 + hz_2^2 \quad (17)$$
$$= c_1 z_1^2 - z_1 z_2 + h\sigma^2.$$

Ensure that Q is positive definite matrix, then

$$\dot{V}_2 \le -z^T Qz - hn|\sigma| \le 0. \quad (18)$$

All these show that

$$|Q| = h(c_1 + hk_1^2) - (hk_1 - \frac{1}{2})^2 = h(c_1 + k_1) - \frac{1}{4}. \quad (19)$$

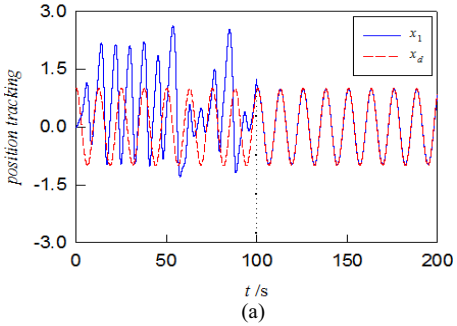
If we need $|Q| \le 0$,we have to get right value of h , c_1 and k_1 .

C. *Numerical Simulations*

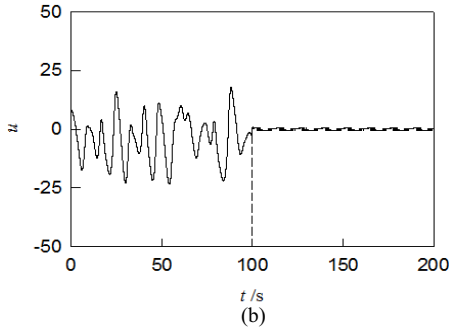
According to the results from section2 to section3, adding controller to equation(4) and setting $k = 2.4269$, and then the model can be rewritten as

$$\begin{cases} \frac{dx_1}{d\tau} = x_2, \\ \frac{dx_2}{d\tau} = -\sin x_1 - 0.4x_2 + 0.2 + 0.02\cos(0.8\tau) \\ \qquad -1.3\cos(1.94\tau)\sin x_1 + u, \end{cases} \quad (20)$$

where $d(t) = 0.02\cos(0.8\tau) - 1.3\cos(1.94\tau)\sin x_1$. The parameters are given $\gamma = 30$, $c_1 = 10$, $k_1 = 10$, $h = 20$. Choosing the position instruction $x_d = \cos(\pi t)$ and $x_d = 0$, respectively, thus the simulation results are shown in Fig.5 and Fig.6. From these four pictures, chaotic motions interfere in the system at first 100 seconds. After that, the controller is added to the system, and then no chaos oscillation phenomenon is observed. Finally, the controlled power system traces the expected target.

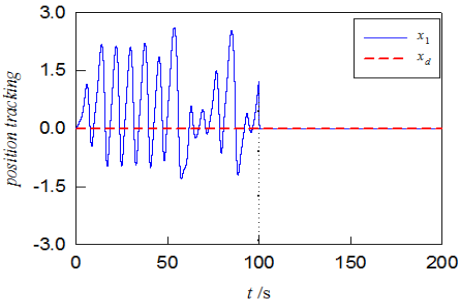


(a)

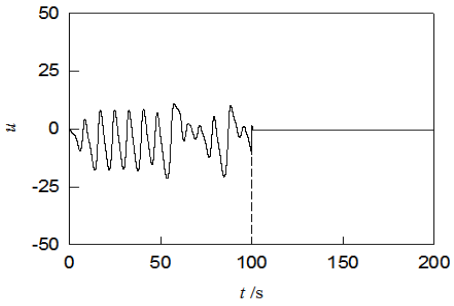


(b)

Figure 5. The adaptive back-stepping sliding mode control process of system based on the position instruction $x_d = \cos(\pi t)$, (a) position tracking (b) control law curve.



(c)



(d)

Figure 6. The adaptive back-stepping sliding mode control process of system based on the position instruction $x_d = 0$, (c) position tracking (d) control law curve.

IV. CONCLUSION

In this paper, a dynamical model of power system was mainly discussed which impacted by some uncertain disturbances, such as load disturbance and electromagnetic power disturbance. The dynamical behavior of the power system was investigated by using the Lyapunov function, phase maps and Poincaré maps of the attractors. We have proven that when parameters were given by some particular values, the chaos oscillation will seriously interfere with our model. Therefore, the back-stepping sliding mode controller was designed to make the system output trajectory gradually follow the target and no longer do random movement. Simulation results demonstrated that the controller have excellent response and tracking performance, chaotic oscillation of this system was suppressed in a short time.

ACKNOWLEDGMENT

This work is supported by Jiangsu Province Ordinary University Graduate Students Scientific Research Innovation Projects of China under Grant: KYLX_0722.

REFERENCES

[1] Lu Qiang, Mei Shengwei and Sun Yuanzhang, "Power system nonlinear control",Beijing:Tsinghua University Press, 2008,chapter 2.

[2] Wang Yibei, Luo Man, Xiao Yanting,Chen Hougui, "Research on chaos phenomena in power system",PEAM,Wu Han,China,2011.

[3] Yu Yixin,Jia Hongjie,Li Peng,Su Jifeng,Power system instability and chaos,14th PSCC,Sevilla,2002.

[4] Jia Hongjie,Yu Yixing,Li Peng,Torus bifurcation and chaos in power systems, Proceedings of the CSEE, 22(8):6-10,2002.

[5] Song Dunwen,Jiang Suna,Hao Jianhong,Bin Hong,A review of low frequency osillation mechanism in power system based on bifurcation and chaos, East China Electric Power,42(6):1115-1122,2014.

[6] Zhang Weidong,Zhang Weinian,Analysis of parameters for chaotic power systems, Power System Technology, 24(12):17-20,2000.

[7] Dong Shiyong,Bao Hai,Wei Zhe,Calcolations and simulations of the chaotic oscillation threshold in Daul-unit systems ,Proceedings of the CSEE,30(19):58-63,2010.

[8] Ni Junkang,Liu Chongxin,Pang Xia,Fuzzy fast terminal sliding mode controller using an equivalent control for chaotic oscillation power system, Acta Phys.Sin,62(19): 190507,2013.

[9] Min Fuhong,Ma Meiling,Zhai Wei,Wang Enrong, Chaotic control of the interconnected power system based on the relay characteristic function,Acta Phys.Sin,63(5):050504,2014.

[10] Liu Jinkun Matlab simulation for sliding mode control, Beijing:Tsinghua University Press,2012:206-212.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Constructing Kernels for One-class Support Vector Machine

Bin Zhang
School of IoT Engineering
Jiangnan University
Wuxi, China
kayzhb@163.com

Jiagang Zhu
School of IoT Engineering
Jiangnan University
Wuxi, China
zhujg@jiangnan.edu.cn

Haobing Tian
School of IoT Engineering
Jiangnan University
Wuxi, China
howingtian@163.com

***Abstract*—OCSVM (one-class support vector machine) is a variant of SVM which only use positive class sample set in training. Since only positive samples can be used in OCSVM, Fully exploiting and using the features of the training samples is of great significance to improve its classification performance. Thus, two aspects of study on kernels have been done in this paper: first, we propose a kernel constructing method called WFCD (weighted feature-contribution-degree) kernel constructing method, in which a PCA (principal component analysis) is performed to the training samples to obtain a vector set with the dimension being sorted by corresponding eigenvalues and then using this vector set to apply a weighed kernel method to concentrate on the larger eigenvalue dimensions; second, we employ the Fisher kernel in OCSVM to decide whether a kernel constructed based on the training sample set has better performance. Experimental results on UCI standard data sets indicate that our method outperforms the general kernel methods and promotes the classification effect considerably.**

Keywords-One-class SVM*; *Fisher Kernel*; *PCA

I. Introduction

Support vector machine(SVM), motivated by the VapnikCChervonenkis (VC) dimensional theory and the statistical learning theory, is a promising machine learning technique which can solve the two-class classification problem effectively.There are many advantages in SVM such as converging to the global optimum, dimension insensitive, strong generalization ability, etc.

When applying the SVM theory into solving the one-class problem, the one-class SVM is proposed. One-class problems are often encountered in practical application area such as fault diagnosis, face recognition, network anomaly detection and text classification. One-class problem can be described as follows: Only one-class samples can be used during training. References [1,2] proposed an one-class SVM idea in which a hypersphere covering all positive class samples is constructed in nonlinear mapping space and then the classification is performed in terms of whether a test sample is within this hypersphere or not.

Reference [3] indicated that, in one-class SVM, the traditional one-class SVM is sensitive to outliers because the classification hyper-plane depends only on a small proportion of the training samples (i.e. support vectors). So it produces overfitting problem. To solve this problem, reference [5] introduced the rough set theory into one-class support vector machine and proposed rough one-class SVM.

On the other hand, the selection of kernel is always an important research field of SVM and has an important impact on the classification performance. Since there are only positive class samples in the training set, how to make a full use of information contained in the training set to construct a better kernel then becomes one of the critical factors to the OCSVM. Based on the above idea, two aspects of study on kernels have been done in this paper.

First, we proposed a kernel constructing method called WFCD (weighted feature-contribution-degree) kernel constructing method, in which a PCA (principal component analysis) to the training sample set is performed to obtain a vector set with the dimensions being sorted by its eigenvalues and then using this vector set to construct a kernel with larger eigenvalue dimension having more impact on the kernel value. As an example, we use RBF to construct WFCD kernel, the relevant kernel is called λ-RBF. Let X be the given training set, we can get its covariance matrix C and the eigenvector matrices P of C by using PCA, then the vector set $X' = P^T X$. λ-RBF is a linear combination of the RBF terms composed of each dimension of X', and the weight of each RBF term is related to the value of its dimensional eigenvalue, with dimension corresponding to larger eigenvalue playing the more important role in the kernel.

Second, we employed a Fisher kernel in OCSVM to explore whether a kernel constructed based on the training sample set has better performance, for such a kernel may make a better use of information contained in training samples.

The results of experiments on UCI standard data sets showed that, compared with the general kernels, the WFCD kernel and Fisher kernel can effectively improve recognition rate of OCSVM by more than 3 percent points on average.

II. One-class SVM and Relevant Kernels

A. One-class SVM[5]

Given a training set without any class information, one class SVM constructs a decision function that takes the

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.123

value +1 in a small region capturing most of the data points, and -1 elsewhere. The strategy in this technique is to map the input vectors into a high-dimensional feature space corresponding to a kernel and construct a linear decision function in this space to separate the dataset from the origin with maximum margin. Via the freedom to utilize different types of kernel, the linear decision functions in the feature space are equivalent to a variety of nonlinear decision functions in the input space. A parameter $v \in (0,1]$ was introduced into the one class SVM to control the trade-off between the fraction of data points in the region and the generalization ability of the decision function.

Given a training dataset without any class information

$$X = (x_1, x_2, \cdots, x_l), x \in R^d \tag{1}$$

In order to separate the dataset from the origin, we solve the following quadratic programming problem (QPP):

$$\begin{aligned} &\min_{\omega,\xi,\rho} \frac{1}{2}\|\omega\|^2 - \rho + \frac{1}{vl}\sum_{i=1}^{l}\xi_i \\ &s.t. (\omega \cdot \phi(x_i)) \geq \rho - \xi_i, \\ &\xi_i \geq 0, i = 1, 2 \cdots, l. \end{aligned} \tag{2}$$

Where ξ_i is a slack variable. Parameter $\delta > 1$ is a threshold parameter.$v \in (0,1]$ is a parameter chosen a priori.

The solution to this QPP(2) is transformed into its dual problem by the saddle point of the Lagrange function,

$$\begin{aligned} &\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\mathbf{K}(x_i, x_j) \\ &s.t. \sum_{i=1}^{l}\alpha_i = 1, 0 \leq \alpha_i \leq \frac{1}{vl}, i = 1, 2, \cdots, l. \end{aligned} \tag{3}$$

Once the solution $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)^T$ to the QPP(3) has been found, the decision function can be expressed as follows:

$$f(x) = sgn\left(\sum_{i=1}^{l}\alpha_i\mathbf{K}(x_i, x) - \rho\right) \tag{4}$$

We can recover threshold ρ by exploiting that for any such $0 \leq \alpha_i \leq \frac{1}{vl}$, the corresponding pattern x_j satisfies

$$\rho = (\omega \cdot \phi(x_j)) = \sum_{i=1}^{l}\alpha_i\mathbf{K}(x_i, x_j) \tag{5}$$

Where $\mathbf{K}(x_i, x_j)$ is a kernel function that gives the dot product$(\phi(x_i) \cdot \phi(x_j))$ in the higher dimensional space.

B. RBF Kernel

The main problem of Support Vector Machine (SVM) is how to train a linear machine with Margin. It is proposed to solve the linearly inseparable problem in the original space by a kernel which mapping it to a higher dimensional space. In SVM, the similarity of two samples is estimated by their inner product. These inner products are strongly dependent on the selected kernel. Different kernel means a different valuation standard. So in solving practical problems, the choice of the kernel is very important. We often need to construct the kernel function according to the corresponding problem.

RBF kernel is one of the kernels used most frequently. The traditional RBF is as follows:

$$k(x_i, x_j) = e^{-\frac{\|x_i, x_j\|^2}{2\sigma^2}} \tag{6}$$

For the given training dataset X, the corresponding kernel matrix is:

$$\mathbf{K}_{\text{matrix}} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_l) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_l) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_l, x_1) & k(x_l, x_2) & \cdots & k(x_l, x_l) \end{bmatrix} \tag{7}$$

C. Fisher Kernel

Fisher Kernel is not frequently used, but it may be another better choice to use it in OCSVM.

Consider a parametric generative model $P(x|\theta)$ where θ denotes the vector of parameters. The goal is to find a kernel that measures the similarity of two input vectors x and x induced by the generative model [14]. In particular, consider the Fisher score:

$$g(\theta, x) = \nabla_\theta lnP(x|\theta) \tag{8}$$

where $P(x|\theta)$ is a distribution determined by training sample set.

From which the Fisher kernel is defined by

$$k(x, x') = g(\theta, x)^T F^{-1} g(\theta, x') \tag{9}$$

where F is the Fisher information matrix:

$$\mathbf{F} = \mathbf{E}_x\left[g(\theta, x) g(\theta, x)^T\right] \tag{10}$$

The expectation in formula is with respect to x under the distribution $P(x|\theta)$.

In practice, one approach is simply to replace the expectation in the definition of the Fisher information with the sample average, giving

$$\mathbf{F} \simeq \frac{1}{N}\sum_{n=1}^{N} g(\theta, x_n) g(\theta, x_n)^T \tag{11}$$

More simply, we can just omit the Fisher information matrix altogether and use the invariant kernel.

$$k(x, x') = g(\theta, x)^T g(\theta, x') \tag{12}$$

III. WFCD-Based RBF Kernel and Gaussian Distribution Based Fisher Kernel

A. Principal component analysis

PCA is an important method in statistical analysis, which represents the covariance structure with a small part of principal component, and the principal component is a linear combination of the original variables.

Suppose $x = \{x_1, x_2, \cdots, x_n\}$ is a d-dimension random vector, then expectation $p = E(x)$ is estimated by the mean of samples $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)\, C = Cov(x)$, the covariance matrix $C = Cov(x)$ is estimated by the covariance of samples $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$. The covariance matrix can be decomposed into $C = PAP^T$, where $P = (p_1, p_2, \cdots, p_d)$ is the feature vector matrix of C, $A = diag(\lambda_1, \lambda_2, \cdots, \lambda_d)$ is the eigenvalue matrix of C. Then the vector set is defined as $x' = P^T(x - p)$.

B. WFCD based on RBF kernel

To construct WFCD based RBF kernel, we propose the concept contribution degree of principal component. The contribution degree of principal component is $\lambda_i / \sum_{k=1}^{p}\lambda_k, i = 1, 2, \cdots, d$.

To understand the contribution degree, lets see the distributions in different dimensions of the vector set.

Randomly generate two class 15-dimension data point sets X_1-$N(3.5, 1)$ and X_2-$N(1, 1)$, then do PCA to them and get the vector set X'. In Cartesian coordinates, draw the two dimension distributions of the first two dimensions of X', as is shown in left of figure 1. In the same way, draw the two dimension distributions of the last two dimensions of X', as is shown in right of figure 1.

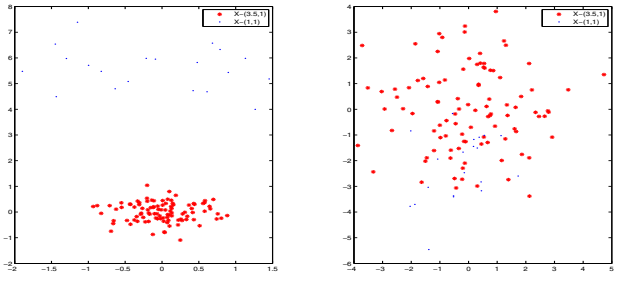


Figure 1. Distribution of different dimensions(left is the first two dimensions and right is the last two dimensions)

As can be seen from the Figures, the distinction degree is higher for the first two dimensions having higher contribution degree, while the distinction degree is lower for the last two dimensions having lower contribution degree. Based on the above characteristic, a WFCD based on RBF kernel is proposed.

For the original training set X and the vector set $X' = \{x_i | x_i \epsilon R^d, i = 1, 2, \cdots, l\}$, the WFCD based on RBF kernel is defined as:

$$k(x_i, x_j) = C_1 e^{\left(x_i^1 - x_j^1\right)} + C_2 e^{\left(x_i^2 - x_j^2\right)} + \cdots + C_d e^{\left(x_i^d - x_j^d\right)} \tag{13}$$

where x_i^d and x_j^d are the d-th dimension of x_i and x_j respectively. Let be i-th eigenvalue of $C = Cov(X)$, with eigenvalues being sorted from large to small, $C_i\,(i = 1, 2, \cdots, d)$ in formula is defined as follows:

$$C_i = \frac{d\lambda_i}{\sum_{j=1}^{d}\lambda_j} \tag{14}$$

Now we explain why formula is a kernel. According to the reference [6], kernels have the following properties:

If $k_1, k_2, \cdots, k_n$ are kernels, then

(1) $\alpha k_i, \alpha \geq 0, i = 1, 2, \cdots, n$ is a kernel;

(2) $k_1 + k_2 + \cdots + k_n$ is a kernel.

Obviously $e^{-\frac{1}{2\sigma^2}\left(x_i^1 - x_j^1\right)^2}, e^{-\frac{1}{2\sigma^2}\left(x_i^2 - x_j^2\right)^2}, \cdots, e^{-\frac{1}{2\sigma^2}\left(x_i^d - x_j^d\right)^2}$ are all RBF kernels in which the input vectors are one dimension.

Because C_i is the constant, according to the property above $C_1 e^{\left(x_i^1 - x_j^1\right)^2}, C_2 e^{\left(x_i^2 - x_j^2\right)^2}, \cdots, C_d e^{\left(x_i^d - x_j^d\right)^2}$ are also kernels. Furthermore, according to property (2) above, $k(x_i, x_j) = C_1 e^{\left(x_i^1 - x_j^1\right)^2} + C_2 e^{\left(x_i^2 - x_j^2\right)^2} + \cdots + C_d e^{\left(x_i^d - x_j^d\right)^2}$ is a kernel which we called λ-RBF kernel.

C. Gaussian distribution based on Fisher kernel

In general case, we can consider that the data conforms to the Gauss model when the data size is large enough. For the sake of convenience, we compute $g(\theta, x)$ with Gaussian single model (GSM). Then

$$\begin{aligned} g(\theta, x) &= \bigtriangledown_\theta lnp(x|\theta) \\ &= [\bigtriangledown_\mu lnp(x|\theta)\ \bigtriangledown_\Sigma lnp(x|\theta)] \end{aligned} \tag{15}$$

Where $\bigtriangledown_\mu lnp(x|\theta) = \Sigma^{-1}(x_t - \mu), \bigtriangledown_\Sigma lnp(x|\theta) = \frac{1}{2}\left(-\Sigma^{-1} + S^T \cdot S\right), S = (x_t - \mu)^T \Sigma^{-1}$.

So we can compute the kernel value with the above formula.

IV. Experiments and Analysis

In order to test the performance of proposed λ-RBF kernel and Fisher kernel in solving the classification problem, we respectively carry out numerical experiments on simulated data and Pen- based Handwritten Digits (PHD) from UCI.

In order to further verify the performances of both λ-RBF kernel and Fisher kernel in classification, we conducted numerical experiments in the PHD UCI database. The basic information of data set is listed in table I below. The experiments consist of 10 kinds of sample data, and there are 16 feature descriptions in each class.

During the experiment we respectively train the classifier using RBF, λ-RBF and fisher kernel each time. Both of the λ-RBF and RBF take the dimensions which contribution rate

Table I
THE BASIC INFORMATION OF PHD

Class	0	1	2	3	4	5	6	7	8	9
Train	780	779	780	719	780	720	720	778	719	719
Test	363	364	364	336	364	335	336	364	336	336

is more than 90%. Parameters v and are selected with five-fold cross-validation. Table II gives the experimental results of fisher kernel, λ-RBF, RBF and non-PCA RBF on PHD data set, where PR denotes the mean reorganization rate of the samples from positive class, NR denotes the mean reorganization rate of the samples from negative class and AR denotes the mean reorganization rate of all samples from both classes.

Table II
PERFORMANCE COMPARISONS BETWEEN FISHER,λ-RBF, RBF AND NON-PCA RBF ON PHD (%)

	FISHER OCSVM			**λ-RBF OCSVM**			**PCA RBF OCSVM**			**non-PCA RBF OCSVM**		
Class	*AR*	*PR*	*NR*	*AR*	*PR*	*NR*	*AR*	*PR*	*NR*	*AR*	*PR*	*NR*
0	97.53	98.08	97.47	96.68	84.3	98.12	95.74	84.02	97.1	94.57	81.82	96.04
1	92.12	90.32	92.33	90.57	91.76	90.43	71.07	79.4	70.1	72.76	83.57	71.51
2	99.2	94.23	99.78	92.97	91.21	93.17	90.85	89.84	90.97	89.77	91.76	89.53
3	98.8	96.5	99.04	96.08	94.35	96.27	92.05	95.83	91.65	93.48	87.09	91.07
4	96.12	92.65	98.77	95.77	72.25	98.5	90.99	79.95	92.28	94.94	84.89	96.11
5	95.19	95.83	95.13	93.42	83.58	94.47	88.19	85.97	88.43	85.16	91.64	84.48
6	98.8	93.75	99.34	96.68	82.44	98.2	92.82	81.85	93.99	92.22	83.93	93.11
7	94.46	92.26	94.71	92.65	72.8	94.96	83.25	85.71	82.96	91.42	84.34	92.25
8	95.79	95.1	95.87	93.17	78.57	94.72	83.25	78.27	83.78	86.36	90.48	85.93
9	93.19	91.61	93.36	94.03	77.68	95.76	85.16	82.14	85.48	88.71	83.33	89.28

From the perspective of prediction accuracy, we can find that the average AR value of λ-RBF OCSVM and fisher kernel OCSVM are up to 94.2% and 96.12% respectively, while the PCA RBF OCSVM and non-PCA RBF OCSVM is 87.34% and 88.94%, so the proposed λ-RBF is better than the other two RBF OCSVM but poorer than fisher kernel OCSVM. For PR, there is certain fluctuation in different test samples with λ-RBF OCSVM, but the performance is better than the other two RBF OCSVM and also poorer than fisher kernel OCSVM. In the aspect of detection for outliers, we can see from the experimental data that λ-RBF OCSVM has a strong stability with a recognition rate of 95.46% that almost the same to fisher kernel OCSVM, while that of the PCA RBF OCSVM and non-PCA RBF OCSVM are 87.67% and 88.93% respectively. From the experiment on UCI we can find that the classification performance of λ-RBF OCSVM is better than the other two RBF but a little poorer than Fisher kernel OCSVM.

V. CONCLUSION

The selection of kernel is more important to OCSVM than to traditional SVM, since only one class samples can be used in training. According to the properties of kernel, we proposed a WFCD based kernel constructing method and apply the method into RBF so that λ-RBF is gotten which is confirmed to have better performance in solving classification problems. At the same time, we studied the Fisher kernel and got the Gaussian based Fisher kernel which is also found to have better performance in OCSVM, if training samples are Gaussian distribution or the similar.

The future work is to study how to extend the method above to other traditional kernels or distributions to get more suitable kernels with better performance in practical uses.

ACKNOWLEDGMENT

The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] D M J, Duin R P W. Support Vector Data Description [J] .Machine Learning, 2004, 54(1): 45-66.

[2] Campbell C, Bennett P. A Linear Programming Approach to Novelty Detection [M]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001.

[3] Wang L, Jia H D, Li J. Training Robust Support Vector Machine with Smooth Ramp Loss in the Primal [J]. Neurocomputing, 2008, 71: 3020-3025.

[4] Lin C-F, Wan g S-D. Fuzzy Support Vector Machine [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.

[5] Yitian Xu , Chunmei Liu. A rough margin-based one class support vector machine. Neural Comput & Applic 2013, 22:1077C1084.

[6] Christopher M. Bishop. Pattern Recognition and Machine Learning [M]. Cambridge: Springer, 2007:291-320.

[7] WANG Lei, YANG Yi-fan, ZHOU Qi-hai. Rough Set based One-class Support Vector Machine[J]. Computer Science. 2009,36(9):242-245.

[8] QIN Yu-ping, WANG Yi, LUN Shu-xian et al. Multi-label Text Classification Algorithm Based on Hyper Ellipsoidal SVM. Computer Science. 2013,40(11A):98-100.

[9] Young-Seon Jeong, In-Ho Kang, Myong-Kee Jeong et al. A New Feature Selection Method for One-Class Classification Problems [J]. IEEE Transactions on Systems, Man, and Cyberneticsłpart c: Applications and Reviews, 2012,42(6):1500-1509.

[10] LI Kai, LU Xiao-xia. A Rough Margin Based Fuzzy Support Vector Machine. Acta Electronica Sinica. 2013,41(6):1183:1187.

[11] Asharaf S, Shevade SK, Narasimha murty M. Rough support vector clustering[J]. Pattern Recogn. 2005, 38(10):1779C1783.

[12] Bicego M, Figueiredo MAT. Soft clustering using weighted one-class support vector machines. Pattern Recogn [J]. 2009, 42(1):27C32.

[13] Imran N. Junejo, Adeel A. Bhutta, Hassan Foroosh. Single-class SVM for dynamic scene modeling [J]. SIViP (2013) 7:45C52.

[14] Shawe-Taylor J,Cfistianini N. Kernel methods for pattern analysis [M]. CambridgeCambridge University Press, 2004.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A High Accuracy Spectral Element Method for Solving Eigenvalue Problems

Weikun Shan
Institute of Software, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
shanweikun66@126.com

Huiyuan Li
Institute of Software, Chinese Academy of Sciences
Beijing, China
huiyuan@iscas.ac.cn

***Abstract*—A triangular spectral element method is proposed and analyzed for the Laplacian eigenvalue problem. The method is based on the Galerkin approximation with generalized Koornwinder polynomials. We detailedly describe the approximation scheme and implementation for solving the Laplacian eigenvalue problem. Numerical experiments also indicate that the triangular spectral element method for solving the eigenvalue problems on convex domain has the "spectral" accuracy, that is, exponential convergence rate.**

Keywords-triangular spectral element method; eigenvalue problem; Koornwinder polynomials; "spectral" accuracy.

I. INTRODUCTION

As is well known, Laplace operator occurs in many important differential equations such as Wave equations and Diffusion equation, etc.. Then solving the Laplacian eigenvalue problem is part of a standard recipe to tackle those equations. Abundant literature is contributed to numerical investigations to Laplacian eigenvalues by various conforming and/or nonconforming finite element methods, including approximation schemes, parallel algorithms and implementations, posterior error estimates, lower and upper bounds estimate, etc.. In contrast, little attention has been paid to the spectral element methods in the numerical approximation of eigenvalue problems. Spectral element methods inherit the high accuracy and convergence rate of the traditional spectral methods, while preserve the flexibility of the low order finite element methods. Evidence shows that spectral element methods enjoy some essential priorities over the traditional spectral method and other low order methods for eigenvalue problems[1, 2, 3].

Spectral element method was first introduced by Patera[4] for Chebyshev expansions, then generalized to the Legendre case by Maday and Patera[5]. The spectral element discretization depends on both the geometric partition and the polynomial degree. It represents a special case of Galerkin methods in which the finite dimensional spaces of the trial/test functions are made of continuous piecewise algebraic polynomials of high degree on the computational domain.

The classic quadrilateral spectral element method exhibits the advantages of using tensorial basis functions and naturally diagonal mass matrices[10,11]. Recently, considerable progress has been made in the triangular spectral element method which is proven to be more flexible for complex domains and for adaptivity. In general, the current approaches dealing with triangular spectral element method can be classified as i) using the Koornwinder-Dubiner polynomials; ii) using non-polynomials functions on triangular domains; iii) using special nodal points as the interpolation points.

The purpose of this paper is to propose a triangular spectral element method and conduct a comprehensive numerical analysis for the following Laplacian eigenvalue problem with homogeneous Dirichlet boundary condition:

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\Omega \subset \mathbb{R}^2$ denotes a bounded domain with Lipschitz boundary $\partial\Omega$. Based on the variational formulation of (1) and the triangulations of Ω, the conforming triangular spectral element approximation scheme is established. To ensure the effectiveness of our approximation scheme, the eigenfunctions are approximated by globally continuous piecewise polynomials of total degree $\leq M$. Special kinds of orthogonal Koornwinder polynomials are then utilized in our triangular spectral element method as the local basis functions on each element such that the corresponding matrix eigenvalue problem can be fast assembled and then be efficiently evaluated by sparse eigen-solvers.

II. THE GENERALIZED JACOBI POLYNOMIALS AND KOORNWINDER POLYNOMIALS

Let Ω be a bounded domain and w be a generic weight function. Denote by $(u,v)_{w,\Omega} = \int_\Omega u(x)v(x)w dx$ and $\|\cdot\|_{w,\Omega}$ the inner product and the norm of $L^2_w(\Omega)$, respectively. Further, we use $H^r_w(\Omega)$ and $H^r_{0,w}(\Omega)$ to denote the usual weighted Sobolev spaces, whose norms and seminorms are denoted by $\|\cdot\|_{r,w,\Omega}$ and $|\cdot|_{r,w,\Omega}$, respectively. In cases where no confusion would arise, w(if $w \equiv 1$) and Ω may be dropped from the notations. The polynomial space $\mathbb{P}_k$ denotes the space of polynomials of total degree $\leq k$. Besides, let $\mathbb{N}$ and $\mathbb{N}_0$ be the collections of the positive integers and non-negative integers, respectively.

A. Generalized Jacobi polynomials

Let $I = (-1,1)$. The hypergeometric representation for the classic Jacobi polynomials $J_n^{\alpha_1,\alpha_2}(x)$, $x \in I, n \in \mathbb{N}_0$ with $\alpha_1, \alpha_2 > -1$ is

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.124

$$J_n^{\alpha_1,\alpha_2}(x) = \binom{n+\alpha_1}{n} {}_2F_1(-n, n+\alpha_1+\alpha_2+1; \alpha_1+1; \frac{1-x}{2}), \tag{2}$$

with $-n-\alpha_1-\alpha_2 \notin \{1,2,\ldots,n\}$. When $\alpha_1, \alpha_2 > -1$, $J_n^{\alpha_1,\alpha_2}$ are mutually orthogonal with respect to the weight function $w^{\alpha_1,\alpha_2} := w^{\alpha_1,\alpha_2}(x) = (1-x)^{\alpha_1}(1+x)^{\alpha_2}$ on I [6,7]. Extending the indices α_1, α_2 to be negative integers, we have obtained the generalized Jacobi polynomials[7]. Of our great interest are the polynomials $J_n^{-1,-1}$. Supplemented for $J_1^{-1,-1}$, and then by (2), we obtain the complete system:

$$\begin{aligned} &J_0^{-1,-1}(x) = 1, \quad J_1^{-1,-1}(x) = x, \\ &J_n^{-1,-1}(x) = \frac{x-1}{2}\frac{x+1}{2} J_{n-2}^{1,1}(x), \ n \geq 2. \end{aligned} \tag{3}$$

Such a supplementation preserves the symmetry properties of the classic Jacobi polynomials:

$$J_n^{\alpha_1,\alpha_2}(x) = (-1)^n J_n^{\alpha_2,\alpha_1}(-x), n \in \mathbb{N}_0, \ \alpha_1, \alpha_2 \geq -1. \tag{4}$$

For more about supplementations of $J_n^{\alpha_1,\alpha_2}$, refer to [8].

B. Koornwinder polynomials

On reference triangle $\hat{T} = \{(\hat{x}, \hat{y}) : 0 < \hat{x}, \hat{y}; \hat{x}+\hat{y} < 1\}$, the standard Koornwinder polynomials $\mathcal{J}_l^{\alpha}(\hat{x}, \hat{y}), \ l \in \mathbb{N}_0^2$ with $\alpha \in (-1, \infty)^3$ can be defined through the Jacobi polynomials:

$$\begin{aligned} \mathcal{J}_l^{\alpha}(\hat{x}, \hat{y}) = (\hat{y}+\hat{x})^{l_1} J_{l_1}^{\alpha_1,\alpha_2}\left(\frac{\hat{y}-\hat{x}}{\hat{y}+\hat{x}}\right) \times \\ J_{l_2}^{2l_1+\alpha_1+\alpha_2+1,\alpha_3}(1-2\hat{x}-2\hat{y}), \quad (\hat{x}, \hat{y}) \in \hat{T}. \end{aligned} \tag{5}$$

The standard Koornwinder polynomials are mutually orthogonal with respect to the weight function $w^{\alpha}(\hat{x}, \hat{y}) = \hat{x}^{\alpha_1}\hat{y}^{\alpha_2}(1-\hat{x}-\hat{y})^{\alpha_3}$and satisfy the following symmetry relations as well:

$$\mathcal{J}_l^{\alpha}(\hat{x}, \hat{y}) = (-1)^{l_1} \mathcal{J}_l^{\alpha_2,\alpha_1,\alpha_3}(\hat{y}, \hat{x}). \tag{6}$$

By virtue of the generalized Jacobi polynomials defined in (3), we extend the Koornwinder polynomials defined in (5) to $\alpha = (-1,-1,-1)$. To obtain uniform formulations for their differentiation relations, we adopt the following normalized representation:

$$\begin{cases} \tilde{\mathcal{J}}_l(\hat{x}, \hat{y}) = \dfrac{2(2l_1-1)(2|l|-1)}{(l_1-1)(|l|+l_1-1)} \mathcal{J}_l^{-1,-1,-1}(\hat{x}, \hat{y}), \quad l_1 \geq 2, \, l_2 \geq 1, \\ \qquad = -\dfrac{2(2l_1-1)(2|l|-1)}{(l_1-1)l_2} \hat{x}\hat{y}(1-\hat{x}-\hat{y}) \mathcal{J}_{l_1-2,l_2-1}^{1,1,1}(\hat{x}, \hat{y}) \\ \tilde{\mathcal{J}}_{l_1,0}(\hat{x}, \hat{y}) = \dfrac{2(2l_1-1)}{l_1-1} \mathcal{J}_{l_1,0}^{-1,-1,-1}(\hat{x}, \hat{y}) \\ \qquad = -\dfrac{2(2l_1-1)}{l_1-1} \hat{x}\hat{y} \mathcal{J}_{l_1-2,0}^{1,1,0}(\hat{x}, \hat{y}), \quad l_1 \geq 2, \\ \tilde{\mathcal{J}}_{0,l_2}(\hat{x}, \hat{y}) = \dfrac{2l_2-1}{l_2-1} \mathcal{J}_{0,l_2}^{-1,-1,-1}(\hat{x}, \hat{y}) \\ \qquad = \dfrac{2l_2-1}{l_2-1}(\hat{x}+\hat{y})(\hat{x}+\hat{y}-1) \mathcal{J}_{0,l_2-2}^{0,0,1}(\hat{x}, \hat{y}), \ l_2 \geq 2, \\ \tilde{\mathcal{J}}_{1,l_2}(\hat{x}, \hat{y}) = \dfrac{2l_2+1}{l_2+1} \mathcal{J}_{1,l_2}^{-1,-1,-1}(\hat{x}, \hat{y}) \\ \qquad = \dfrac{2l_2+1}{l_2}(\hat{x}-\hat{y})(\hat{x}+\hat{y}-1) \mathcal{J}_{0,l_2-1}^{0,0,1}(\hat{x}, \hat{y}), \ l_2 \geq 1, \\ \tilde{\mathcal{J}}_{1,0}(\hat{x}, \hat{y}) = \hat{y}-\hat{x}, \quad \tilde{\mathcal{J}}_{0,1}(\hat{x}, \hat{y}) = -\hat{y}-\hat{x}, \quad \tilde{\mathcal{J}}_{0,0}(\hat{x}, \hat{y}) = 1. \end{cases} \tag{7}$$

III. THE TRIANGULAR SPECTRAL ELEMENT METHOD APPROXIMATION SCHEME AND IMPLEMENTATION

In this section, we devote to the approximation scheme and implementation for the Laplacian eigenvalue problem of the triangular spectral element method.

A. Variational formulations and approximation scheme

The variational form of the Laplacian eigenvalue problem (1) can be presented as follows: to find $\lambda \in \mathbb{R}$ and $u \in H_0^1(\Omega)$ with $u \neq 0$ such that

$$a(u, v) := (\nabla u, \nabla v) = \lambda(u, v) \quad \forall v \in H_0^1(\Omega). \tag{8}$$

It is obvious that $a(\cdot,\cdot)$ is continuous on $H_0^1(\Omega) \times H_0^1(\Omega)$. Besides, $a(\cdot,\cdot)$ is coercive, i.e.,

$$a(u, u) \geq c\|u\|_1^2 \quad \forall u \in H_0^1(\Omega), \tag{9}$$

with certain constant $c > 0$.

Let $\mathcal{T} = \{T_i\}$ be a triangular partition of Ω. We assume that the partition satisfies the following regularities,

- Each element T_i is *spectrally admissible* in the sense that there is a bijective mapping F_i of class C^{∞}, which maps $\overline{\hat{T}}$ onto $\overline{T_i}$. In our paper, we only consider the case that every T_i has only straight sides such that each F_i is an affine mapping. The case in which T_i have curvilinear sides is out of our consideration.
- $\mathcal{T}$ is *regular*, that is, $\bar{T}_i \cap \bar{T}_j, \ i \neq j$ is either empty or it consists of a vertex or an entire side of T_i.
- $\mathcal{T}$ is *shape regular* which means that there exists a constant κ, independent of any $T_i \in \mathcal{T}$, such that

$$\sup_{T_i \in \mathcal{T}} \frac{h_{T_i}}{\rho_{T_i}} \leq \kappa < \infty, \tag{10}$$

where h_{T_i}, ρ_{T_i} denote the diameters of the circumcircle and the incircle of the element T_i, respectively.

Denote by M the degree of the polynomial utilized by the spectral element method on each element $T_i \in \mathcal{T}$. Let $\delta = \delta(h, M)$ be the discretization parameter with the mesh size $h = \max_{T_i \in \mathcal{T}} h_{T_i}$. We now define the approximation space $V_\delta \subset H_0^1(\Omega)$ as follows:

$$V_\delta = \{v \in H_0^1(\Omega) : v|_{T_m} \circ F_m \in \mathbb{P}_M(\hat{T}), T_m \in \mathcal{T}\}. \tag{11}$$

Then the Galerkin spectral element approximation scheme for the Laplacian eigenvalue problem (8) reads as: to find $\lambda_\delta \in \mathbb{R}$ and $0 \neq u_\delta \in V_\delta$ such that

$$a(u_\delta, v) = \lambda_\delta(u_\delta, v) \quad \forall v \in V_\delta. \tag{12}$$

B. Implementation of the spectral element method

Now, in order to implement (12) by the triangular spectral element method, we make use of the Koornwinder polynomials $\tilde{\mathcal{J}}_l$ to construct the basis functions which are divided into vertex modes, edge modes and interior modes. The vertex mode only has a magnitude at one vertex and is zero at other vertices; the edge mode has magnitudes on one edge and is zero on other edges and vertices; while the interior mode is identically zero on all edges and vertices.

- Vertex mode:

$$\begin{aligned}(0,0):&\quad \phi_{0,0}(\hat{x},\hat{y})=1-\hat{x}-\hat{y},\\ (1,0):&\quad \phi_{1,0}(\hat{x},\hat{y})=\hat{x},\\ (0,1):&\quad \phi_{0,1}(\hat{x},\hat{y})=\hat{y},\end{aligned}\tag{13}$$

- Edge mode:

$$\begin{aligned}\hat{x}+\hat{y}=1:&\quad \phi_{l_1,0}(\hat{x},\hat{y})=\bar{\mathcal{J}}_{p,0}(\hat{x},\hat{y}), \qquad l_1\geq 2,\\ \hat{y}=0:&\quad \phi_{0,l_2}(\hat{x},\hat{y})=\tilde{\mathcal{J}}_{0,l_2}(\hat{x},\hat{y})+\tilde{\mathcal{J}}_{1,l_2-1}(\hat{x},\hat{y}), \quad l_2\geq 2,\\ \hat{x}=0:&\quad \phi_{1,l_2-1}(\hat{x},\hat{y})=\tilde{\mathcal{J}}_{0,l_2}(\hat{x},\hat{y})-\tilde{\mathcal{J}}_{1,l_2-1}(\hat{x},\hat{y}), \quad l_2\geq 2,\end{aligned}\tag{14}$$

- Interior mode:

$$\phi_{l_1,l_2}(\hat{x},\hat{y})=\tilde{\mathcal{J}}_{l_1,l_2}(\hat{x},\hat{y}),\ l_1\geq 2,\ l_2\geq 1.\tag{15}$$

We list the function expansion and the derivative expansions of $\tilde{\mathcal{J}}_l$ which can be readily derived from the appendix part in [7]. For $|l|=l_1+l_2\geq 2$, it holds that

$$\begin{aligned}\tilde{\mathcal{J}}_{l_1,l_2}=&\frac{(|l|+l_1)(|l|+l_1+1)}{2|l|(2|l|+1)}\mathcal{J}^{0,0,0}_{l_1,l_2}-\frac{(l_2+1)(l_2+2)}{2|l|(2|l|+1)}\mathcal{J}^{0,0,0}_{l_1-2,l_2+2}\\ &-\frac{(|l|+l_1)(|l|-3l_1-1)}{(2|l|+1)(2|l|-2)}\mathcal{J}^{0,0,0}_{l_1,l_2-1}+\frac{(|l|+3l_1-4)(l_2+1)}{(2|l|+1)(2|l|-2)}\mathcal{J}^{0,0,0}_{l_1-2,l_2+1}\\ &-\frac{(l_2-1)(|l|+3l_1)}{2|l|(2|l|-3)}\mathcal{J}^{0,0,0}_{l_1,l_2-2}+\frac{(|l|+l_1-2)(|l|-3l_1+3)}{2|l|(2|l|-3)}\mathcal{J}^{0,0,0}_{l_1-2,l_2}\\ &+\frac{(l_2-1)(l_2-2)}{(2|l|-3)(2|l|-2)}\mathcal{J}^{0,0,0}_{l_1,l_2-3}-\frac{(|l|+l_1-3)(|l|+l_1-2)}{(2|l|-3)(2|l|-2)}\mathcal{J}^{0,0,0}_{l_1-2,l_2-1},\end{aligned}\tag{16}$$

$$\begin{aligned}\partial_{\hat{x}}\tilde{\mathcal{J}}_{l_1,l_2}=&-(|l|+l_1)\mathcal{J}^{0,0,0}_{l_1,l_2-1}-(2l_1-1)\mathcal{J}^{0,0,0}_{l_1-1,l_2}\\ &+(l_2+1)\mathcal{J}^{0,0,0}_{l_1-2,l_2+1}+(l_2-1)\mathcal{J}^{0,0,0}_{l_1,l_2-2}\\ &-(2l_1-1)\mathcal{J}^{0,0,0}_{l_1-1,l_2-1}-(|l|+l_1-2)\mathcal{J}^{0,0,0}_{l_1-2,l_2},\end{aligned}\tag{17}$$

$$\begin{aligned}\partial_{\hat{y}}\tilde{\mathcal{J}}_{l_1,l_2}=&-(|l|+l_1)\mathcal{J}^{0,0,0}_{l_1,l_2-1}+(2l_1-1)\mathcal{J}^{0,0,0}_{l_1-1,l_2}\\ &+(l_2+1)\mathcal{J}^{0,0,0}_{l_1-2,l_2+1}+(l_2-1)\mathcal{J}^{0,0,0}_{l_1,l_2-2}\\ &+(2l_1-1)\mathcal{J}^{0,0,0}_{l_1-1,l_2-1}-(|l|+l_1-2)\mathcal{J}^{0,0,0}_{l_1-2,l_2}.\end{aligned}\tag{18}$$

Hereafter we use the convention that $\mathcal{J}^{\alpha}_{l_1,l_2}=0$ for $l_1<0$ and/or $l_2<0$.

Let Γ be a part of the boundary of $\hat{T}$, which is either an empty set or constituted by one, two or three sides of $\hat{T}$. Define $\mathbb{P}^{\Gamma}_k(\hat{T})=\{v\in\mathbb{P}_k(\hat{T}):v|_{\Gamma}=0\}$. Then, set

$$\begin{aligned}V^{(m)}_M&=\{v\circ F_m^{-1}: v\in\mathbb{P}^{\Gamma}_M(\hat{T}),\ \Gamma=F_m^{-1}(\partial\Omega\cap T_m)\}\\ &=\{\phi_l\circ F_m^{-1}: l\in\Lambda_m\},\end{aligned}\tag{19}$$

where $\Lambda_m=\{l\in\mathbb{N}_0^2:\phi_l\circ F_m^{-1}\in V^{(m)}_M\}$. In such a way, the approximation space V_δ can be written as

$$V_\delta=\{v\in H_1(\Omega):v|_{T_m}\in V^{(m)}_M,\forall T_m\in\mathcal{T}\}.\tag{20}$$

Finally, define the barycentric coordinates $\tau=(\hat{x},\hat{y},1-\hat{x}-\hat{y})$, and then the local stiff and mass matrices on element T_m can be expressed as:

$$\begin{aligned}B^{(m)}=&[2S_m(\phi_k,\phi_j)_{\hat{T}}],\\ A^{(m)}=&\frac{h_2^2+h_3^2-h_1^2}{4S_m}A^{2,3}+\frac{h_3^2+h_1^2-h_2^2}{4S_m}A^{3,1}\\ &+\frac{h_1^2+h_2^2-h_3^2}{4S_m}A^{1,2},\\ A^{\ell,l}=&[((\partial_{\tau_\ell}-\partial_{\tau_l})\phi_k,(\partial_{\tau_\ell}-\partial_{\tau_l})\phi_j)_{\hat{T}}],\end{aligned}\tag{21}$$

where S_m,h_1,h_2,h_3 denote the area and the length of the three sides of element T_m. In the end, assembling the local matrices, we get the global stiff and mass matrices A, B, and then the algebraic eigenvalue system can be expressed by

$$AU=\lambda_\delta BU,\tag{22}$$

where all the matrix entries can be analytically evaluated through (16), (17) and (18) without any ultilization of quadrature rules.

For our Laplacian eigenvalue problem (1), the deduced eigen-system (22) has positive definite stiff matrix and mass matrix, thus can be efficiently solved by algebraic eigenvalue package such as ARPACK.

IV. Numerical Results

In this section, we carry out some numerical computation for the implementation of the triangular spectral element approximation for the Laplacian eigenvalues on both the square and the L-shaped domain.

A. Square domain

We first consider the Laplacian eigenvalue problem (1) on the square $\Omega=[-1,1]^2$ by the triangular spectral element method. It is well known that the Laplacian eigenvalues of problem (1) on the square have the following representation:

$$\lambda=\frac{(k_1^2+k_2^2)\pi^2}{4},k_1,k_2\geq 1.\tag{23}$$

Denote by λ_i and $\lambda_{i\delta}$ the eigenvalues of (8) and (12) sorted in ascending order. The relative errors $|\lambda_i-\lambda_{i\delta}|/\lambda_i$ versus M/h of the 5 smallest eigenvalues are then plotted in Fig.1(right), Fig.2(right) and Fig.3(right) for the corresponding meshes with mesh size $h=2,h=1$ and $h=1/2$ on the left, respectively.

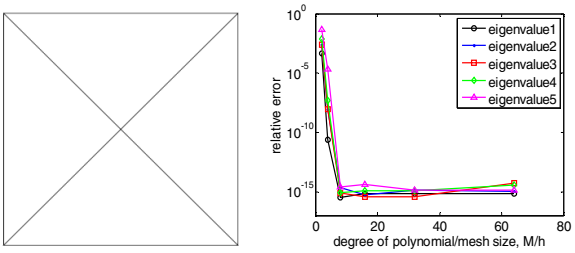


Fig.1. Left: uniform triangulation on Ω with mesh size $h=2$; Right: relative errors of the five smallest eigenvalues corresponding to the mesh on the left.

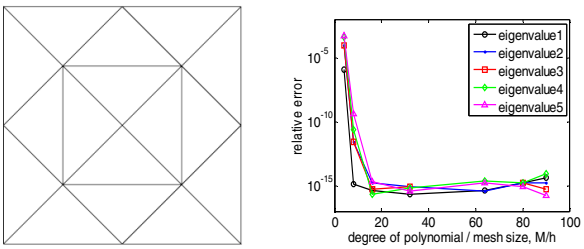


Fig.2. Left: uniform triangulation on Ω with mesh size $h=1$; Right: relative errors of the five smallest eigenvalues corresponding to the mesh on the left.

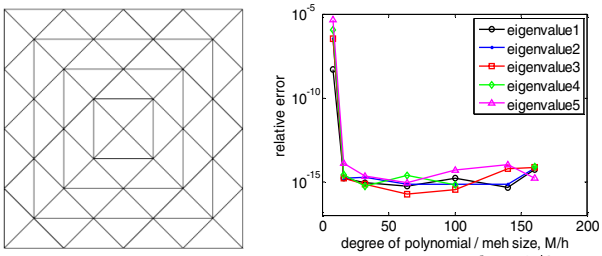


Fig.3. Left: uniform triangulation on Ω with mesh size $h = 1/2$; Right: relative errors of the five smallest eigenvalues corresponding to the mesh on the left.

As indicated in Fig.1, Fig.2 and Fig.3, for a fixed mesh size h, the computational eigenvalues have reached exponential convergence rate. For comparison, the 5 smallest eigenvalues $\lambda_i^e (i = 1, \cdots, 5)$ computed by the linear finite element method (FEM) are tabulated in Table I-Table IV as well as the relative errors and the convergence rate. We can observe that the discrete eigenvalues of the linear finite element method converge from above to the exact ones as the mesh size h tends to zero and the convergence rate is approximately $r \approx 2$.

TABLE I. computation of the first eigenvalue of Laplacian

Mesh size	$\lambda_1^e \times \frac{4}{\pi^2}$	$\frac{\lvert\lambda_1^e - \lambda_1\rvert}{\lambda_1}$	*rate*
h=1/2	2.098934776481971	0.04946738824099	/
h=1/4	2.025331529636071	0.01266576481804	1.9655435
h=1/8	2.006394465437555	0.00319723271878	1.9860385
h=1/16	2.001603964046409	0.00080198202321	1.9951819
h=1/32	2.000401415002385	0.00020070750119	1.9984754
h=1/64	2.000100385733632	0.00005019286682	1.9995403

TABLE II. computation of the second and third eigenvalue of Laplacian

Mesh size	$\lambda_{2,3}^e \times \frac{4}{\pi^2}$	$\frac{\lvert\lambda_{2,3}^e - \lambda_{2,3}\rvert}{\lambda_{2,3}}$	*rate*
h=1/2	5.522143605551108	0.10442872111022	/
h=1/4	5.135641457408183	0.02712829148164	1.9446485
h=1/8	5.034246288351649	0.00684925767033	1.9857786
h=1/16	5.008584110347835	0.00171682206957	1.9962071
h=1/32	5.002147475502045	0.00042949510041	1.9990269
h=1/64	5.000536960168887	0.00010739203378	1.9997547

a. The second and third eigenvalues are repeated eigenvalues.

TABLE III. computation of the forth eigenvalue of Laplacian

Mesh size	$\lambda_4^e \times \frac{4}{\pi^2}$	$\frac{\lvert\lambda_4^e - \lambda_4\rvert}{\lambda_4}$	*rate*
h=1/2	9.618809107463319	0.20235113843292	/
h=1/4	8.411811839672161	0.05147647995902	1.9748757
h=1/8	8.102862436588204	0.01285780457353	2.0012691
h=1/16	8.025705651319273	0.00321320641491	2.0005588
h=1/32	8.006425748211616	0.00080321852645	2.0001492
h=1/64	8.001606394988814	0.00020079937360	2.0000378

TABLE IV. computation of the fifth eigenvalue of Laplacian

Mesh size	$\lambda_5^e \times \frac{4}{\pi^2}$	$\frac{\lvert\lambda_5^e - \lambda_5\rvert}{\lambda_5}$	*rate*
h=1/2	12.451997407229216	0.24519974072292	/
h=1/4	10.639318311067063	0.06393183110671	1.9393511
h=1/8	10.163711996389033	0.01637119963890	1.9653744
h=1/16	10.041281815388578	0.00412818153886	1.9875817
h=1/32	10.010347919786183	0.00103479197862	1.9961656
h=1/64	10.002588988730519	0.00025889887305	1.9988802

B. *L-shaped domain.*

In our second example, we consider the Laplacian eigenvalue problem (1) on the L-shaped domain $\Omega = [-1, 1]^2 \backslash [0, 1]^2$, which is obtained by removing the top right quadrant of the square. The L-shaped domain has one reentrant corner, which may induce the corner singularity of the eigenfunctions and thus admits only a limited convergence rate for some eigenvalues. Convergence results of the triangular spectral element method with uniform triangulation of mesh size $h = 1/2$(Fig.4 left) are reported in Fig.4(right). Here, the reference eigenvalues are obtained by the triangular spectral element method using the geometric mesh(as shown in Fig.5) of level 15 with $M = 50$, since it also guarantees an exponential convergence, as indicated by Babuška[9].

Fig.4(right) shows the relative errors of the 5 smallest eigenvalues, among which an asymptotic convergence rate around 2.6 is observed for the first and fifth eigenvalues. For the second and forth eigenvalues, the asymptotic convergence rate is about 5.2. Moreover, exponential convergence rate is achieved for the third eigenvalue. But anyway, the accuracy of the triangular spectral element method is higher than the linear finite element method.

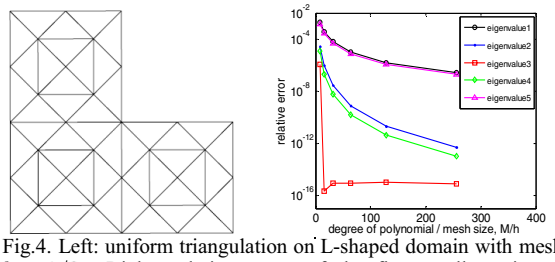


Fig.4. Left: uniform triangulation on L-shaped domain with mesh size $h = 1/2$; Right: relative errors of the five smallest eigenvalues corresponding to the mesh on the left.

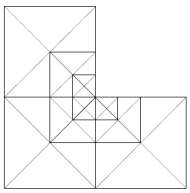


Fig.5. Geometric mesh of level 3.

V. CONCLUSION

In summary, we have proposed the triangular spectral element method for the Laplacian eigenvalue problem with homogeneous Dirichlet boundary conditions based on the generalized Koornwinder polynomials, then emphasis has been set on the comprehensive numerical analysis on the triangular spectral element method. Compared with the classic finite element method, the triangular spectral element method can achieve exponential convergence rate in the case that the computational domain is convex. However, if the computational domain has reentrant corners, a limited convergence rate will be observed for some eigenvalues. Despite this, the spectral element method is not inferior to the finite element method, and has good applicability as well.

The use of the orthogonal Koornwinder polynomials greatly simplifies the calculation such that, for the algebraic eigenvalue system, all the matrix entries can be analytically evaluated through the expansion relations (16), (17) and (18) without any ultilization of quadrature rules. Meanwhile, the sparsity of the matrices is also good. On the other hand, in the process of solving eigenvalue problems by spectral element method, the program can be well paralleled in view of the independence of the calculation on each element. Moreover, taking into account the characteristics of the spectral element methods, one can further design a more efficient algorithm. And most of all, in many cases, the spectral element method can achieve exponential convergence rate, thus we can obtain desired convergence results without the need of quite dense mesh and high order basis functions. Given the above, we conclude that the spectral element method for solving eigenvalue problems is a high accuracy numerical method which is suitable for large-scale parallel computing.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (No. 91130014, 11471312 and 91430216).

REFERENCES

[1] J. Weideman and L. Trefethen, The eigenvalues of second-order spectral differentiation matrices, SIAM Journal on Numerical Analysis, 25 (1988), pp. 1279–1298.

[2] J. P. Boyd, Chebyshev and Fourier spectral methods, Courier Dover Publications, 2001.

[3] Z. Zhang, How many numerical eigenvalues can we trust?, Journal of Scientific Computing, (2014), pp. 1–12.

[4] A. T. Patera, A spectral element method for fluid dynamics: laminar flow in a channel expansion, Journal of computational Physics, 54 (1984), pp. 468–488.

[5] Y. Maday and A. T. Patera, Spectral element methods for the incompressible Navier-Stokes equations, in IN: State-of-the-art surveys on computational mechanics (A90-47176 21-64). New York, American Society of Mechanical Engineers, 1989, p. 71-143. Research supported by DARPA., vol. 1, 1989, pp. 71–143.

[6] B.-Y. Guo, J. Shen, and L.-L. Wang, Optimal spectral-Galerkin methods using generalized Jacobi polynomials, Journal of Scientific Computing, 27 (2006), pp. 305–322.

[7] H. Li and J. Shen, Optimal error estimates in Jacobi-weighted Sobolev spaces for polynomial approximations on the triangle, Mathematics of Computation, 79 (2010), pp. 1621–1646.

[8] H. Li and Y. Xu, Spectral approximation on the unit ball, SIAM Journal on Numerical Analysis, 52 (2014), pp. 2647–2675.

[9] I. Babuška, B. Q. Guo, and E. P. Stephan, On the exponential convergence of the h–p version for boundary element Galerkin methods on polygons, Mathematical methods in the applied sciences, 12 (1990), pp. 413–427.

[10] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, Spectral Methods: Fundamentals in Single Domains, Scientific Computation, Springer-Verlag, Berlin, 2006.

[11] G. E. Karniadakis and S. J. Sherwin, Spectral/hp element methods for computational fluid dynamics, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, second ed., 2005.

The Multi-class SVM is Applied in Transformer Fault Diagnosis

Liping Qu
Electric & Information Engineering School
Beihua University
Jilin, China
E-mail: goudan1990@126.com

Haohan Zhou
Electric & Information Engineering School
Beihua University
Jilin, China
E-mail: zhouhaohan1313@126.com

Abstract—Transformer fault forecast plays an important role in the safe and stable operation of power system. So it is important to detect the incipient faults of transformer as early as possible. In this study, the support vector machine (SVM) is introduced to analyze and diagnosis the transformer fault. According to the accumulation fault data, the SVM forecast model take the RBF as the kernel function and utilize the best pattern to cope with data for reducing imbalance. In order to prove the SVM method efficacious and accuracy, we also make the diagnosis with traditional three ratio method experimental. The results of the final experimental indicate that SVM can make higher diagnosis accuracy and have excellently generalization ability.

Key word—Support Vector Machine; Transformer Fault; Fault Diagnosis; Experimental

I. INTRODUCTION

Transformer is one of the most important electrical equipment in the power system, and it is related directly to the security and stability of the power system. Through predicting the transformer fault types, we can find the potential failure in time and then make timely processing. The transformer fault diagnosis of dissolve gas analysis (DGA)[1] is the current research hotspot and the most convenient for the technology of transformer diagnosis. Because different transformer faults correspond to produce gases' composition are different. In recently years, in order to get better performance of fault diagnosis, a variety of diagnostic methods have been studying, such as: the characteristic gas method, IEC three ratio method[3], four ratio method[4], graphic method, etc, and the artificial intelligence[5,6] also research based on gas analysis.

According to traditional methods, those are need infinity fault samples to gain feasible diagnosis results. However, sometimes, there is no much reality data. Support Vector Machine(SVM) is a class of supervised learning algorithms introduced by Vapnik firstly. And it is a kind of building on the basis of statistical learning machine learning method. It can solve the problem of neural network algorithm of high dimension and local minima problem, and be used for transformer fault diagnosis very well. The fault diagnosis classification is based on structural risk minimization principle [9].And its is very powerful for the problem with limited samples. In this paper ,in view of transformer DGA, we utilize multi-class SVM in one-versus-one method to diagnosis transformer more fault types.

II. PRINCIPLE OF THE FAULT DIAGNOSIS

A. Principle of transformer fault diagnosis

In order to ensure the stable operation of power system, transformer will work for a long time if there is no something wrong in it. Due to thermal breakdown and electrical failure will cause the decomposition of transformer oil, insulating paper and dissolve in transformer oil. The common gases are as follows [10]:

Hydrogen(H_2);methane(CH_4);ethane(C_2H_6);ethylene(C_2H_4);
acetylene (C_2H_2);carbon monoxide(CO);carbon dioxide (CO_2).

The first five kinds of gases (*hydrogen (H2), methane (CH4), ethane (C2H6), ethylene (C2H4), acetylene (C2H2)*) were often used to analysis. Daily operation and maintenance will regularly and irregularly use gas chromatographic analyzer to detect all kinds of dissolved gases. Considering a large amount of accumulation and summarization, researchers infer and acknowledge that different failure parts would been caused by the different gas content. Thus, researcher and profession can identify the transformer fault types by the gas content. This article also base that and adopts the way of SVM classification of fault type classification method to diagnosis the transformer fault.

B. Brief introduction to SVM

SVM was originally designed by binary classification problem [6], and the basic idea can be illustrated in figure 1.1.The triangle and circle are the two points, which represent the different data in transformer fault. L2, L3 are the support straight line, and the distance between the two straight lines is $2/\|\omega\|$.The most important is to find the right line for target classification.

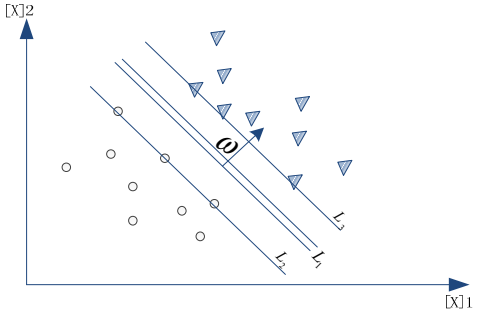


Fig.1 Binary classification model diagram

In above, the SVM linear classification decision function is as follows:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.125

$$\min \quad \frac{1}{2}\omega^T \cdot \omega + C\sum_i^n \xi_i \tag{1}$$

Subject to:

$$\begin{cases} y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, i = 1,2,...n. \\ \xi_i \geq 0, (i = 1,2,...n) \end{cases} \tag{2}$$

Here, C, as constant，is the penalty factor. To solve above problem, the solution method of the optimization problem can be given through Lagrange functional:

$$L(\omega, b, \xi, \alpha, \beta) = \frac{1}{2}(\omega^T \cdot \omega) + \mathrm{C}\sum_i^n \xi_i - \sum_{i=1}^n \alpha_i \{y_i[\omega \cdot \Phi(x_i) + b] - 1 + \xi_i\} - \sum_{i=1}^n \beta_i \xi_i \tag{3}$$

Here, α_i, β_i are the square operator and $\alpha_i \geq 0, \beta_i \geq 0$. Quadratic programming problems are as follows:

$$\max \quad W(\alpha, \beta) = \max_{\alpha,\beta}\{\min_{\omega,b,\xi} L(\omega, b, \alpha, \beta)\} \tag{4}$$

Therefore，we will be get：

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{i=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{5}$$

Above all, subject to: $0 \leq \alpha_i \leq C, \sum_{i=1}^n \alpha_i y_i = 0$

Then, b^* & α_i^* will be got and we can make the decision function and it likes the following form:

$$f(x) = \mathrm{sgn}(y(\mathrm{x})) \tag{6}$$

Here $y(x) = (\omega \cdot x) + b = \sum_{i=1}^n \alpha_i^* y_i (x_i \cdot x) + b^*$

III. MULTI-CLASS SVM IS APPLIED IN TRANSFORMER FAULT DIAGNOSIS

Transformer faults are divided into hot fault and electrical fault, and the common failure types are the follows [10]:

(1)Low temperature overheat;(2)Middle temperature overheat;(3)High temperature overheat;(4)Partial discharge;(5)Low energy discharge;(6)Low energy discharge and overheating;(7)Arc discharge;(8)Arc discharge and overheating.

Moreover, there is a state of normal work and set it to types 9. As we all know, the original SVM is only a kind of two kinds of linear classifier. Thus, we need to find a way to solve more than a kind of classifier. For the purpose of multi-class, various different binary classification methods are implemented[7], such as 'one-against-all', 'one-against-one', 'binary tree' etc. According to analysis the transformer fault type, we choose the method of one-against-one. Its' structure model is as follows:

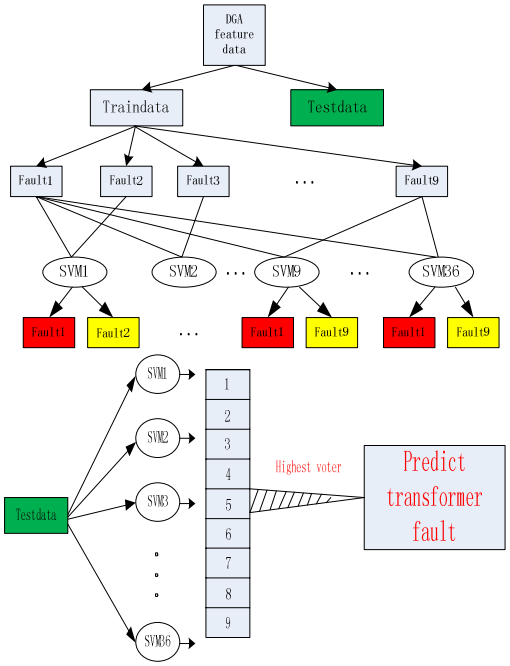


Fig.2 Diagnostic model of power transformer based on mutil-class SVM

In the case of complex nonlinear transformer fault diagnosis, this article introduces the kernel function which maps the original data into a high-dimensional.The construction and selection of kernel function are very important to SVM classical. There are four basic kernels[7]:

（1）Linear $K(x_i, x_j) = (x_i \cdot x_j')$

（2）Polynomial $K(x_i, x_j) = (r \cdot (x_i \cdot x_j') + 1)^d$

（3）RBF $K(x_i, x_j) = \exp(-\gamma \cdot \| x_i - x_j \|^2)$

（4）Sigmoid $K(x_i, x_j) = \tanh(r \cdot (x_i \cdot x_j') + r0)$

According to the kernel function, the decision function can be written as:

$$f(x) = \mathrm{sgn}[\sum_{i=1}^n \alpha_i^* y_i K(x_i, x) + b^*] \tag{7}$$

In order to avoid to be affected by the imbalance of the types data, contents of these diagnostic gases are preprocessed through a special data processing. The preprocessed forms are as follows:

Pattern1： $x_s^* = \dfrac{x_s}{\max_{i=1}(x_i)}$

Pattern 2： $x_s^* = \dfrac{x_s - x_{\min}}{x_{\max} - x_{\min}}$

Pattern 3： $x_s^* = \dfrac{x_s}{\sum_{i=1}^n (x_i)}$

Here $x_{\max}$ and $x_{\min}$ represents the maximum value and the minimum value in the data set respectively; x_i(i=1 ,2, 3, 4, 5)is

the concentration of the raw gas data; x_s are the actual data and x_s^* are preprocessed data. The different patterns of the transformer fault diagnosis accuracy will be different.

III. SVM MODEL AND EXPERIMENTAL RESULT ANALYSIS

There are 223 groups of transformer fault types collecting from different transformer substation and have been examined actual transformer fault. The transformer fault data are divided into training-data and testing-data. In this paper, we choose five kinds of gas as input features. Sample classification of various experimental fault types shown in the table 1 below:

TABLE1 Sample classification of various experimental fault types

Number	Fault Types	Actual test fault	Training data(127)	Testing data(96)
1	Low temperature overheat (<150)	Line overheat	10	8
	Low temperature overheat (150~300℃)	Tap-changer poor contact, lead screw loosening or bad welding joint, eddy current fever, iron core magnetic flux leakage, short circuit and multipoint earth		
2	Middle temperature overheat (300~700℃)		17	16
3	High temperature overheat (>700℃)		50	30
4	Partial discharge	Moisture, gas discharge	4	3
5	Low energy discharge	Spark discharge between the floating potential components	10	8
6	Low energy discharge and overheating	Tap and oil discharge gap between different parts	4	4
7	Arc discharge	Between interterm, interlayer, alternate with, lead or to discharge, tap-changer arc discharge caused by circulation, etc	19	19
8	Arc discharge and overheating		7	4
9	Safe	Normal	6	4

Firstly we cope with the fault data in above three patterns. According to experiment such as follow table 2, we can find that the pattern3 will be high accuracy. Therefore, we choose the patterns3 as coping with the fault data.

TABLE2 Comparison of diagnosis result among different patterns

Pattern	Testing-data	Accuracy_SVM
Pattern 0	96	34.375%
Pattern 1	96	89.5833%
Pattern 2	96	89.5833%
Pattern 3	96	95.8333%

Then, because the transformer fault diagnosis is the nonlinear, this article introduces the kernel function which maps the original data into a high-dimensional. And since the radial basis function (RBF) only needs to set one parameter, we choose the RBF kernel function.

$$K(x_i, x_j) = \exp(-\gamma \cdot \| x_i - x_j \|^2) \quad (8)$$

Then, we can use those clues to make the diagnosis model for the transformer fault prediction.

In order to illustrate the mulit-SVM diagnosis methods valid, this article uses the same transformer fault data to predict the fault types with the tradition three ratio method according to the international electro-technical commission(IEC)[10] .The simulation of the two method results are as follows:

(1)The IEC transformer diagnosis:

Accuracy =88.54% (85/96)

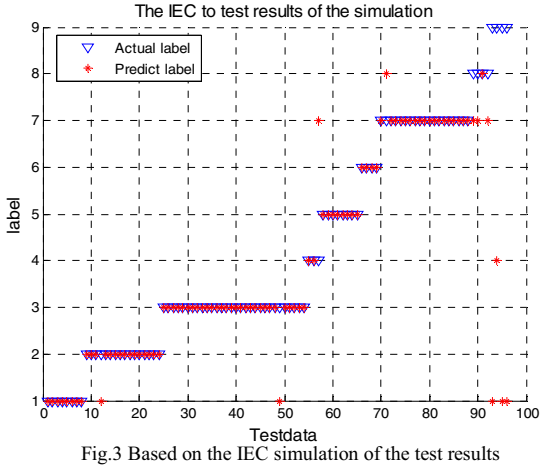


Fig.3 Based on the IEC simulation of the test results

(2)The SVM transformer diagnosis:

Accuracy =94.79%(91/96)

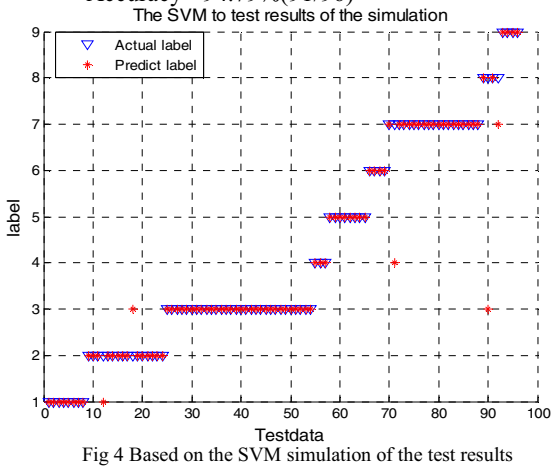


Fig 4 Based on the SVM simulation of the test results

The results show that this article method can solve the classification of transformer fault diagnosis and predict the transformer fault types effective. Furthermore, from the figure3 and figure4, comparing with the three ratio method of fault diagnosis accuracy (88.54%), this paper method has a higher accuracy (94.79%) in transformer fault diagnosis.

V. CONCLUSION

In this study, the mutil-class SVM model has been set up and to detect the transformer fault. From experimental results, this method greatly improved the correct diagnosis rate and the accuracy is 94.79%. Thus, we can conclude that the multi-class SVM can realize the improvement of diagnostic accuracy. Furthermore, it is suitable for being used in the development of online fault diagnosis system for power transformers and will accelerate the development of the intelligent power system.

In the future research, considering the parameter optimization importance for SVM, our work will choose other intelligence algorithm to search the best parameters. And also, in order to get excellently transformer fault diagnosis model and improve accuracy greatly, some further studies will be performed to combine other methods with SVM.

REFERENCE

[1] A Review of Faults Detectable by Gas-in-Oil Analysis in Transformers [J]. IEEE Electrical Insulation Magazine.2002, 18(3):8-17.

[2] Zhu Deiheng, Yan zhang, Tan Kexiong. Electrical equipment status inspection and fault diagnosis technology [M].Beijing: China Electric Power Press, 2009.03:244-302.

[3] Chen Qiming, Tang Wen. Comparative study on three kinds of transformer fault diagnosis method[J].Power System Technology, 2006,10(30):423-425.

[4] Zou jian, Lu Jing,Zhou Xiaofang. Four ratio method in the application of the transformer overheating fault judgment[J].Transformer, 2011.48(10):66-67.

[5] Castro A R G, Miranda V. Knowledge discovery in neural networks with application to transformer failure diagnosis [J].IEEE Transaction on Power Systems, 2005,20(2):717-724.

[6] Vapnik V N.The nature of statistical learning theory[M]. New York: Springer,2000.

[7] Souahlia, Seifeddine;Bacha, Khmais;Chaari, Abdelkader. SVM-based decision for power transformers fault diagnosis using Rogers and Doernenburg ratios DGA.10th Internation Multi-Conference on Systems,Signals&Devices,2013.

[8] Mehta, Amit Kumar,Sharma, R.N.,Chauhan, Sushil,Saho, Satyabrata.Transformer diagnostics under dissolved gas analysis using Support Vector Machine.2013 International Conference on Power,Energy and Control(ICPEC),2013.

[9] Mehta, Amit Kumar;Sharma, R.N.;Chauhan, Sushil;Saho, Satyabrata.Transformer diagnostics under dissolved gas analysis using Support Vector Machine. Power, Energy and Control (ICPEC), 2013 International Conference on,2013.

[10] The state economic and trade commission of the People's Republic of China. DL/T722—2000 Guide dissolved gas in transformer oil, and judgment[S].Beijing: China electric power press,2001.

[11] Wang Xiaochuang, Shi Feng,Yu Lei, Li Yang. Matlab Neural Network 43 Case Analysis[M]. Beijing: Beijing university of aeronautics and astronautics press, 2014:102-179.

[12] Maksim Lapin,Matthias Hein,Bernt Schiele.Learning using privileged information:SVM+and weighted SVM[J].Neural Networks,2014,53: 95-108.

[13] Amit Kumar Mehta,R.N.Sharma,Sushil Chauhan.Transformer Diagnostics under dissolved gas analysis using support vector machine.2013 International Conference on Power,Energy and Control(ICPEC),2013.

[14] M.Bigdeli, M.Vakilian, E.Rahimpour. Transformer winding faults classification based on transfer function analysis by support vector machine.The Institution of Engineering and technology, 2012:268-276.

[15] R. Zhang,J.MA.An improved SVM method P-SVM for classification of remotely sensed data[J]. International Journal of Remote Sensing, 2008,29(20): 6029- 6036.

[16] Eslam Pourbasheer;Siavash Riahi;Mohammad Reza Ganjali;Parviz Norouzi.Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity[J]. European Journal of Medicinal Chemistry,2009,44(12): 5023-5028.

[17] Keskes, H;Braham, A .DAG SVM and pitch synchronous wavelet transform for induction motor diagnosis. Power Electronics, Machines and Drives, 7th IET International Conference on,2014.

A Modified K-means Algorithm based RBF Neural Network and Its Application in Time Series Modelling

Yiping Jiao, Yu Shen, Shumin Fei
School of Automation
Southeast University
Nanjing, China
E-mail: jiaoyiping@163.com

***Abstract*—In this paper, a modified K-means based RBFNN is discussed. To improve the performance of RBFNN, an initial cluster centers (ICCs) selection strategy is proposed for K-means algorithm. The algorithm takes breadth preferred subset of samples as ICCs to cover the sample space using greedy strategy. The results shows that the proposed algorithm can improve the performance of RBFNN remarkably in chaotic time series modelling.**

Keywords- K-means Algorithm; Initial Cluster Centers; RBF Neural Network; Chaotic Time Series

I. INTRODUCTION

Clustering algorithm is widely applied in machine learning [1], which is of value in both theoretical research and actual practice. K-means approach, belongs to partition kind in clustering, is one of the most common used algorithm. To find a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized, K-means approach takes a greedy strategy which generate new partition by assigning each pattern to its closet cluster center and compute new cluster centers in turn [2].

Radical Basis Function (RBF) was firstly introduced to Artificial Neural Network (ANN) by Broomhead [3]. RBF Neural Network (RBFNN) has good properties such as having local response and being trained quickly. The training approach mainly include clustering based method [4] and gradient based method [5]. Clustering based method is the most classic one, which using unsupervised learning (K-means for example) towards samples inputs to determine the centers of hidden nodes, and using Least Minimum Square (LMS) method to determine the outputs weight vector [6].

The disadvantages of K-means are mainly being sensitive to initial cluster centers (ICCs) selection, and the necessity of assigning cluster number K. Many articles discuss the optimization of ICCs like [7]. However, the application of intelligent algorithm or calculating distances of each pair nodes make the K-means approach less simplicity. Affinity Propagation (AP) [8] algorithm needn't give K explicitly, and the performance is excellent, however, the calculation of similarity of pairs makes it low efficiency while faced with large scale problem.

This paper proposed a method to obtain a group of optimized ICCs for K-means, according to a predefined distance, which could be assigned by user using priori knowledge or by trial and error. Cluster number K needn't to be assigned explicitly here. The application in chaotic time series modelling indicates that it can improve the performance of RBFNN significantly.

II. THE MODIFIED K-MEANS BASED RBFNN

The traditional K-manes algorithm is sensitive to initial cluster centers, and often achieve a local optimal solution, which depends on the initial locations of cluster centers [9], this brings much uncertainty to clustering, and can leads to bad results in following steps.

The ICCs selection algorithm is based on an intuitive idea, which deem that a disperse group of nodes is better than completely randomly selected ones as ICCs. The dispersed ICCs made the clustering at a "nearly been well partitioned" state. Moreover, the distribution could better fit all the samples in global. Then the ICCs were used to initialize K-means. The final centers within K-means approach were assigned to the hidden nodes centers in RBFNN. RBFNN can be trained using LMS approach. The structure of the algorithm is showed as Figure.1.

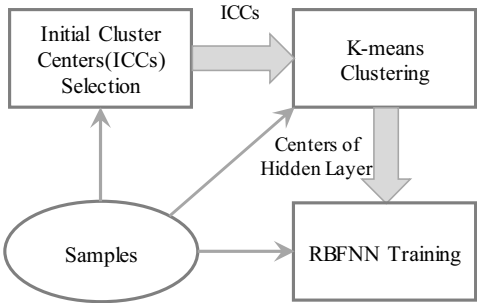


Figure 1. Example of a ONE-COLUMN figure caption.

A. Initial Cluster Centers (ICCs) Selection Algorithm

In this part, the predefined distance d determine the selected centers. The algorithm traverse all samples, calculate the distances between the current sample and the centers already been selected at each step. If all existing centers are far enough to the current sample, the current one is treated as a new center.

The algorithm can be described as Algorithm 1.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.126

Algorithm 1: Initial Cluster Centers Selection

Input: Samples set $S=\{s_1, s_2, ..., s_N\}$, predefined distance d

Output: Selected Initial Cluster Centers C, cluster number K

(1) Initialize: Let $C=\varnothing$, $i=1$

(2) If $C=\varnothing$ or $\min\left(dist\left(C, s_i\right)\right) > d$, then $C = C \cup s_i$

(3) If $i < N$ goto (2), else $K = |C|$

In the algorithm, $dist\left(C, s_i\right)$ denotes the distances between c_j and s_i:

$$dist\left(C, s_i\right) = \left\{d_{ij} \mid d_{ij} = \left\|c_j - s_i\right\|, c_j \in C\right\}$$

and $|C|$ is the elements number in C. The algorithm will selected a group of centers in the given samples, which could cover the samples space roughly. And it is obvious that the set C has following properties, which can be demonstrated as Figure.2.

- $\forall c_j, c_k \in C, j \neq k$, holds $\left\|c_j - c_k\right\| > d$;
- $\forall s_i \in S, \exists c_j \in C$, holds $\left\|c_j - s_i\right\| \leq d$.

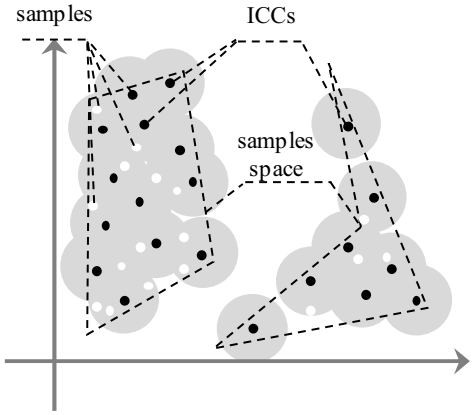


Figure 2. ICCs Selection Results.

B. K-means based RBFNN Training

After initial cluster centers C is obtained, a K-means based RBFNN can be trained by following steps, whose structure is showed as Figure.3.

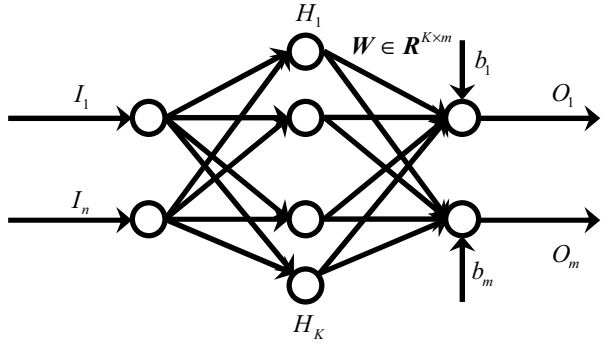


Figure 3. Structure of RBFNN

Algorithm 2: K-means based RBFNN Training

Input: Initial cluster centers C, samples $S=\{s_1, s_2, ..., s_N\}$

Output: Trained RBFNN (Hidden layer centers $\boldsymbol{H}$, spread constants $\boldsymbol{\delta}$, weight matrix $\boldsymbol{W}$)

(1) K-means clustering using C as initial cluster centers. Obtain final cluster centers $\boldsymbol{H}=\{h_1, h_2, ..., h_K\}$.

(2) Calculate spread constants $\delta_j = \kappa d_j$, where κ is the overlap coefficient. d_j is the minimum distance between c_j and other centers, namely $d_j = \min_k \left\|c_j - c_k\right\|$.

(3) $\boldsymbol{H}=\{h_1, h_2, ..., h_K\}$ is set as central vectors of hidden layer nodes in RBFNN.

(4) Calculate hidden nodes output matrix $\hat{\boldsymbol{H}} = \left[h_{ij}\right]$, where $h_{ij} = \Phi_j\left(\left\|s_i - c_j\right\|\right)$ denotes the output of hidden node j while input is sample s_i.

(5) Obtain weight matrix $\boldsymbol{W} = \hat{\boldsymbol{H}}^{+} y$, where $\hat{\boldsymbol{H}}^{+}$ is the pseudo inverse of $\hat{\boldsymbol{H}}$ and y is corresponding dependent variable of samples.

In the above, activation function $\Phi_j(\bullet)$ is radical basis function, for example the Gaussian function $\Phi_j(t) = e^{-\frac{t^2}{\delta_j^2}}$.

III. EXAMPLES

The basis of chaotic time series analysis is phase space reconstruction, whose basic principle derived from delay embedding theorem [10, 11]. For a single variable time series $\{x_n\}$, the phase space reconstruction can be done as:

$$\boldsymbol{x}_n = \left(x_n, x_{n-\tau}, ..., x_{n-(m-1)\tau}\right)$$

and the time series can be modeled as

$$\boldsymbol{x}_{n+1} = \boldsymbol{G}\left(\boldsymbol{x}_n\right)$$

in the formed m-dimension state space.

To illustrate the performance of the algorithm, the modified RBFNN is tested on benchmark time series modelling problem. Samples were taken from time series, and divided into train set and test set. The sum of square error (SSE) of test set were used to evaluate the performance of RBFNN quantificationally.

A. Mackey-Glass Chaotic Time Series [12]

Mackey-Glass (MG) time series is a kind of chaotic time series generated by

$$\frac{dx(t)}{dt} = \frac{\alpha x(t-\tau)}{1 + x^{\gamma}(t-\tau)} - \beta x(t)$$

where $\alpha = 0.2, \beta = 0.1, \gamma = 10$ and τ is adjustable parameter taken as $\tau = 17$.

Take the integration step as $h = 0.01$ to generate MG series, the RBFNN is build to predict $x(t+85)$ using $x(t-18)$, $x(t-12)$, $x(t-6)$ and $x(t)$. The MG time series, and the SSE of RBFNN is showed as Figure 4 and Figure 5.

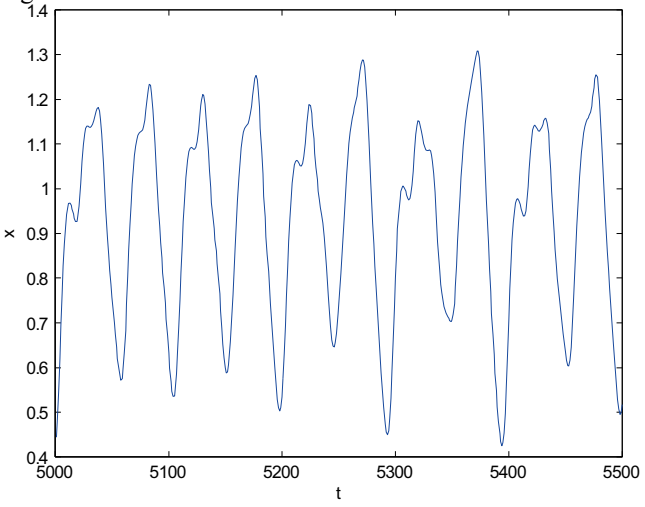


Figure 4. MG Time Series

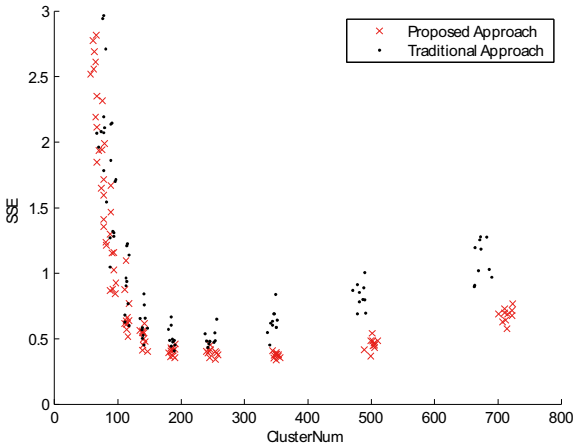


Figure 5. Results of MG Time Series RBFNN Model

In Figure 5, each instance presents the test set SSE of the corresponding RBFNN. The algorithm was executed repeatedly under different cluster number. The cross instances stand for the proposed enhanced k-means based RBFNN, and the dot instances stand for traditional k-means based RBFNN, in which the ICCs were selected randomly from the train set. It can be seen that the proposed approach shows higher performance than the traditional ones. And it also shows trends in decreasing K-means clustering calculation in practice.

B. Lorenz System [13]

The Lorenz system is generated by:

$$\begin{cases} \dfrac{dx}{dt} = \sigma(y - x) \\ \dfrac{dy}{dt} = x(r - z) - y \\ \dfrac{dz}{dt} = xy - bz \end{cases}$$

where parameter $\sigma = 10$, $r = 28$, $b = 8/3$ and the initial condition is $x_0 = 15.34$, $y_0 = 13.68$, $z_0 = 37.91$.

The integration step is also take as $h = 0.01$. Assuming only variable x is observable, the RBFNN is built to predict $x(n+16)$ using $x(n-8)$, $x(n-4)$, and $x(n)$. The Lorenz system and the results of RBFNN is showed as Figure 6 and Figure 7.

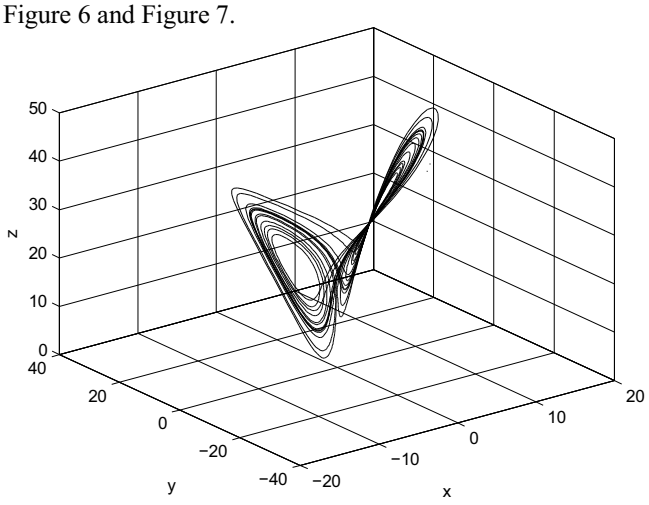


Figure 6. Lorenz System

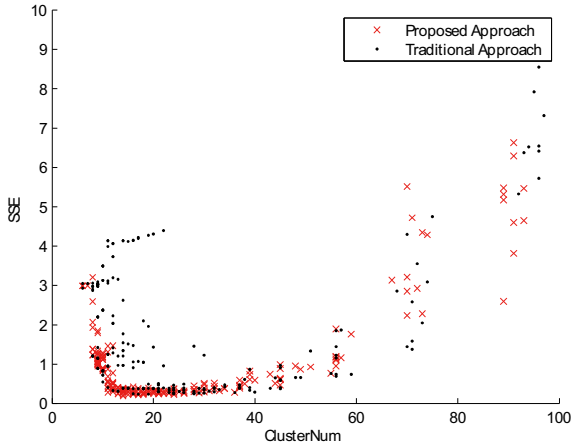


Figure 7. Results of Lorenz System RBFNN Model

The meaning of Figure 7 is similar to Figure 5. It can be seen that the proposed algorithm still show advantages over the traditional one. The best case in this problem is reached by the proposed algorithm at $k^* = 14, SSE^* = 0.1824$.

While the traditional one reach $SSE^*' = 0.2271$ at $k^*' = 17$.

IV. CONCLUSION

This paper proposed an enhanced K-means clustering algorithm, and it can be applied into RBFNN training. The main idea of the algorithm is to select a group of optimized initial cluster centers, which could cover the samples space uniformly. The examples in time series modelling benchmark shows that the proposed algorithm can improve the performance of RBFNN significantly.

Except for ICCs selection strategy, other variant K-means approach, or other clustering approach may lead to different performance while being integrated into RBFNN, which is worth further research.

REFERENCES

[1] Bishop, Christopher M. Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006.

[2] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

[3] BROOMHEAD, DS. "Multivariable functional interpolation and adaptive networks." Complex Systems 2 (1988): 321-355.

[4] Moody, John, and Christian J. Darken. "Fast learning in networks of locally-tuned processing units." Neural computation 1.2 (1989): 281-294.

[5] Hykin, S. "Neural Networks: A Comprehensive Foundation. Printice-Hall." Inc., New Jersey (1999).

[6] Wei, H. K. "The theory and method of neural network structure design."National Defence Industry Press, Beijing (2005).

[7] LIU, Qiang, and Jing-hui WU. "Optimizing initial cluster center of K-means algorithm." Information Technology 2 (2009): 71-73.

[8] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." science 315.5814 (2007): 972-976.

[9] Chen, S. "Nonlinear time series modelling and prediction using Gaussian RBF networks with enhanced cllustering and RLS learning." Electronics letters 31.2 (1995): 117-118.

[10] Takens, Floris. On the numerical determination of the dimension of an attractor. Springer Berlin Heidelberg, 1985.

[11] Mañé, Ricardo. "On the dimension of the compact invariant sets of certain non-linear maps." Dynamical systems and turbulence, Warwick 1980. Springer Berlin Heidelberg, 1981. 230-242.

[12] Crowder, R. Scott. "Predicting the Mackey-Glass time series with cascade-correlation learning." Connectionist Models: Proceedings of the 1990 Summer School. San Mateo, CA, 1990.

[13] Haiyan, Wang, and Lu Shan. "Nonlinear time series analysis and its application." (2006): 13.

Based on rough sets and the associated analysis of KNN text classification research

Guo Aizhang
Qilu University of Technology
Jinan250353, China
gaz@qlu.edu.cn

Yang Tao
Qilu University of Technology
Jinan250353, China
yangtao8812@163.com

***Abstract*—With the rapid development of network information technology, the text is as a basic information carrier and begins to present exponential growth. The existing text classification methods haven't got information from the vast amounts of information resources timely and accurately. In order to solve the problem, the paper puts forward a new method about text categorization. It is a KNN algorithm based on rough set and correlation analysis. Firstly, we introduce the concept of rough set. In the training set of text vector space, we divide all kinds of text vector spaces into certain and uncertain areas. For certain areas, we can directly judge its category. For uncertain areas, we determine the type of text vector through KNN text classification algorithm based on correlation analysis. Experimental results show that the KNN text classification algorithm based on rough sets and the associated analysis have greatly improved the efficiency and accuracy of text categorization. It can meet the requirements of processing large amounts of text data.**

Keywords- text classification; k-NearestNeighbo; Correlation analysis; The rough set

I. INTRODUCTION

With the rapid development of network information technology, text classification technology has become a major focus of modern information processing. Now, frequent text classification algorithms are: Bayesian[1], support vector machine (SVM) [2], neural network [3], decision number [4], the K - Nearest neighbor [4] and so on. K Nearest neighbor (on K Neighbor, KNN) method has been applied widely in the existing text classification methods. it is a nonparametric classification techniques. It is very effective on pattern recognition based on the statistical. It can achieve higher classification accuracy for the unknown and non-normal distribution. It's advantage is robust, concept is very clear [1].But there are also some obvious disadvantages:

First, when it identifies the category of text which is to be classified, it needs to calculate the similarity of all the samples in the training sample sets. So with the increase of training samples, classification performance will drop soon.

Second, KNN algorithm must specify the k value. Now, there are no better methods to determine the neighbors number for text classification. How to select k value is important to category identification. If k is too big or too small, this will affect the accuracy of text classification.

Third, When KNN algorithm calculate text similarity, it doesn't consider the relationship between characteristic words in texts.

Because time complexity and k value of the traditional KNN text classification method is hard to determine, in order to solve this question, this paper proposes a KNN text classification algorithm based on rough set and correlation analysis, It can improve the efficiency and accuracy of text classification.

II. RESEARCH OF KNN TEXT CLASSIFICATION ALGORITHM

A. Traditional KNN text classification algorithm

Cover and Hart put forward KNN method 1968 firstly. It is a mature algorithm in theory. The basic idea of this algorithm is: According to the traditional vector space model, text content has been converted into a weighted feature vectors in the feature space. For example, keywords sets $T=\{t_1,t_2,t_3...t_n\}$,text sets $D=\{d_1,d_2,d_3...d_n\}$,then, we use a vector to represent a document. That is $d_j=(t_1,w_1;t_2,w_2;...t_n,w_n)$ .For a testing text, we calculate each text's similarity in all training sample sets .Find the k most similar texts, According to weighted distance and test texts' category. Specific algorithm steps are as follows:

1) for a test text, according to the formation words to get test text vector.

2) Calculate text similarity of each text in the test text and the training set. The formula are as follows:

a) cosine:

$$\mathrm{Sim}(\mathrm{d_i},\mathrm{d_j})=\frac{\cos\theta=\sum_{k=1}^{n}\mathrm{w_{ik}}\times\mathrm{w_{jk}}}{\sqrt{\sum_{k=1}^{n}\mathrm{w_{ik}}^2\times\sum_{k=1}^{n}\mathrm{w_{jk}}^2}} \quad (1)$$

Equation (1) represents a similarity of text which is relative to another text. The greater the similarity is, the higher the degree of correlation for two texts is.

b) Vector produce:

$$Sim(d_i,d_j)=d_i\times d_j=\sum_{k=1}^{n}w_{ik}\times w_{jk} \quad (2)$$

c) Euclidean distance:

$$\mathrm{D}(\mathrm{d_i},\mathrm{d_j})=\sqrt{\frac{1}{\mathrm{N}}\left(\sum_{k=1}^{n}(\mathrm{w_{ik}}-\mathrm{w_{jk}})^2\right)} \quad (3)$$

3) According to the text similarity, find k texts which are similar to test text in the training text sets.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.127

4) Calculate the weight of each class successively in the test text of k-nearest neighbors.

5) Compare the weight class and assign the text to the highest weight category.

B. Improved KNN text classification algorithm

When KNN text classification algorithm is to determine category of text which is to be classified, it needs to calculate similarity of all training sample sets. Then, we choose k samples whose similarity is higher of the text to be necessary to calculate the similarity with the training sample set of all samples. When dealing with high-dimensional text vector or large text data, with increase of the number of training samples, the time and space complexity will also increase, text classification efficiency will decrease. For these shortcomings, paper [6] proposed some improved algorithms .Firstly, the algorithm uses Fuzzy ART to cluster each type of sample in training sample set. It will reduce the amount of data about training sample set .It improves the computational speed of the algorithm and keeps the prediction accuracy. So the algorithm is suitable for large data sets situation. Paper [7] introduces the rough set theory to the text classification. It uses concept of upper and lower approximation to depict the distribution of various types of training samples. It calculates the approximate range of up and down in the training process. According to distribution positions of text vector in the sample space, improved algorithm can determine the position of the text, reduce search range of K-nearest neighbor. Paper [8] is based on paper [7], It divides text vector space into core areas and mix area. It improves the traditional KNN algorithm membership function. For different text area, it adopts different classification scheme for classification. It improves classification accuracy based on Paper [7], based on the same time, the efficiency of classification has been greatly improved.

When KNN algorithm calculates text similarity, it does not consider the relationship between each word text character words. In order to solve this problem, Paper [9] uses characteristics analysis to measure the contribution of the classification. Its character is that polymerization text vector puts the related characteristic words as feature items. It can instead of traditional methods of characteristic words corresponding to a one-dimensional vector. Not only does this reduce the dimension of the vector, but also it strengthens the contribution of text classification feature items. Paper [10] put forward an improved KNN method based on association analysis. Firstly, extracting frequent feature sets and relative text from different training texts, then, we analyze relationship analysis results of total training texts and determine fairish k value. We select k nearest neighbors value in training samples whose category is known by us. Finally, we determine category of unknown text according to neighbor category. Compared with the traditional text classification method based on KNN, improved method can determine k value better and reduce time complexity.

Most improved KNN text classification algorithms only solve one or two questions about text classification method. Although the effect is obvious, but there are still some drawbacks. Paper [7] doesn't have a good solution for upper approximate space. Paper [10] can improve the classification efficiency. But the algorithm treats all sample sets of certain category as a single transaction. It increases the time complexity of the algorithm. In addition, all samples of the class don't only belong to the class. This can lead that final frequent item sets are incorrectly. Therefore, author combined with previous research and proposed KNN text classification algorithm based on rough sets and association analysis. Experimental results show that the improved method can improve the classification speed of the text and can't affect the accuracy of the original algorithm. It can meet the needs of the massive text data classification.

III. KNN TEXT CLASSIFICATION BASED ON ROUGH SETS AND ASSOCIATION ANALYSIS

A. Description of text vector space distribution based on the Rough Set

This paper thinks that spatial region of each class in text classification is vague and uncertain. Rough set is a mathematical tool which represents uncertainty and incompleteness .Paper uses rough set method to describe distribution of text vector space. We introduce up and down similar concept.

Definition 1 In the text vector space, if there exists a subspace, all text vectors of the subspace only belongs to a category C, then we call the maximum subspace C is lower approximation space, written as R(C).

Definition 2 (upper approximation space) In the text vector space, if there exists a subspace, all text vectors of category C belong to the subspace, then we call the minimum subspace C is upper approximation space, written as R (C).

From the above definition we know, lower approximation space R (C) of class C_i is the largest subspace which only contains a class C_i ; Upper approximation space R (C) of class C_i is a minimum subspace which contains the class C_i .

We assume that vector sets of all texts are M. It distributes of the sample space V of m-dimension, the text vector space has n categories. They respectively represent $C_1, C_2, C_3 ... C_n$;O(C_i) represents the center vectors of class C_i . $K(O(C_i), r)$ represents a hyper sphere space . Its radius is r and its center is O(C_i). The text vectors which are in it are $D(K(O(Ci),r))$.

Many experimental data shows that space distribution of the vector texts in M is as following :

- Most of C_i vectors gather around O(C_i). Generally speaking, distance of vectors in other categories is near than that of O(C_i) (the Euclidean distance, the below is same).
- With the decrease of r, percent rate of text vectors which belong to category C_i in $D(K(O(C_i), r))$ will become high.

We make use of this law and use the following method to calculate all kinds of approximate value of upper and lower approximation space range.

We assume that sample vector space is V, the number of vectors which are nearest to c_i are m_i. The number of vectors which belong to C_i are C_i . We assume that $t = \frac{n_i}{m_i}$, when value of m_i is very small, t is equal to 1, With the increase of m_i, it will occur t <1. We need to write down the number and position m_i. We find out the max value r between O(C_i) and the m_i vectors. We can know through text vectors space distribution, The vast majority of vectors in $D(K(O(C_i),r))$ belong to C_i , $D(K(O(C_i),r))$ is equal to lower similar space of C_i approximately. With the increase of m_i, t tends to be zero . We need to write down number and position of m_i. According to law of text vector space distribution, we can know that text vectors of C_i are in C_i . $D(K(O(C_i),r))$ is similar to class upper approximate space of C_i .

B. Text Classification Description based on association analysis

In this paper, we propose text classification description based on association analysis, the basic idea is as follows: Suppose that class text C_i has n text, denoted as $\{d_1, d_2, d_3 ... d_n\}$ respectively, there are x characteristics in each text word t, where x represents number of feature words which each text contains. Words are not included the characteristics weights of each feature word. Each text represents $d_i = t_1, t_2, t_3 ... t_x$. Now we treat each text of C_i as a single transaction (transaction), we regard characteristics of each text as data sets of each affairs. Item sets of each affairs are characteristics collection of a certain text. We set the minimum support. We use Apriori algorithm to get all sets which meet the minimum support threshold in class C_i . The maximum frequent item set is $Ti = \{t_1, t_2, t_3 ... t_m\}$. We can use maximum frequent item sets to reduce dimension.

Then we need to carry on sum operation to weights of feature words which belong to maximum frequent itemsets in text vectors. We use descending order. the biggest weights and the corresponding text category should be unknown classification text belongs.

C. Improved algorithm based on rough sets associated analysis

This paper presents a KNN text classification algorithm based on rough sets and associated analysis. The algorithm combines advantages of the rough set KNN algorithm and association algorithm KNN algorithm. It improve and optimize them. It makes up shortcomings of traditional KNN algorithms. Improved algorithm has two stages:

1) According to the text vector space of rough sets about rough set of upper and lower approximation space, entire text vector space can be divided into two rough sets parts: one is directly clear in the text space of the space for your text category, it is called to identified region, the other is the text space which does not directly determine, it is called the uncertainty.

This paper uses different classification methods based on two different text space.

a) When text vector which to be classified is in determining area, we can directly determine that the classification text belongs to this category. This can narrow the range of uncertainty texts which fall upper approximation space and reducing number of the required text. This can improve classification efficiency.

b) When the text vector which is to be classified is in the uncertainty area, According to definition of upper approximate space, we cannot directly determine the category which it belongs to, Therefore, we need to identify it by the text of association analysis .then we begin the second stage.

2) Classify the text vectors of uncertain region by text classification algorithm of text vector .Specific steps are as follows:

a) In the text vector space ,number of text vector which is in uncertain region is m, Vector of each text represents as $d1, d2, d3 ... dm$. Categories of text vectors are $C1, C2 ... Cn$. For any text vector of uncertain region di , It can be seen as a single transaction. We can treat character item which contains text vector di as data item of each transaction, Each text vector represents become a transaction set of one transaction; For each text vector of uncertain region in category Ci , We can convert it into form of transaction sets and form a collection of category Ci of transaction data. We set the minimum support and use Apriori algorithm to get all items of Ci class text All meet the minimum support threshold set all of the items obtained ci class text maximal frequent itemsets $Ti = t1, t2, t3 ... tx$.

b) For text vector di which is to be classified , we assume it is in uncertain area of category Ch, Cj, C_k, C_l . We also need to calculate total character weight which Ch, Cj, C_k, C_l contains. Assume that weight sum is Ch, Cj, C_k, C_l respectively. The results are in descending order. Category of high weight sum is category which text vector di. if there is ch=cj that is the case that weights sum are equal. Because the feature items which maximal frequent itemsets text contains can represent text properties of this category. The category that we need is one who has the most maximum frequent items characteristic of category ch, cj.

IV. EXPERIMENTAL VALIDATION AND ANALYSIS

We use Web crawler to get the finance, technology, education, cars and real estate in five categories of news from sina news. They are as samples of this experiment, each category were grabbed 2500, news The total is 12,500 documents where each class takes 2000, a total of 10,000 is as training set; the remaining 500 in each

category, a total of 2500 is as a test set. It is used to test classification efficient after algorithm improved. Table 1 shows the distribution of the experimental data in the training set and test set.

TABLE I. CLASSIFICATION DISTRIBUTION OF THE EXPERIMENT IN THE TRAINING SET AND TEST SET.

	Finance and economics	Finance and economics	Education	Car	Estate
Training set	2000	2001	2000	2000	2000
Test Set	500	501	500	500	500

The proposed algorithm compares with KNN text classification algorithm based on rough set, KNN text classification algorithm based on association analysis and traditional KNN text classification algorithm in the classification accuracy and classification accuracy. The results were as follows:

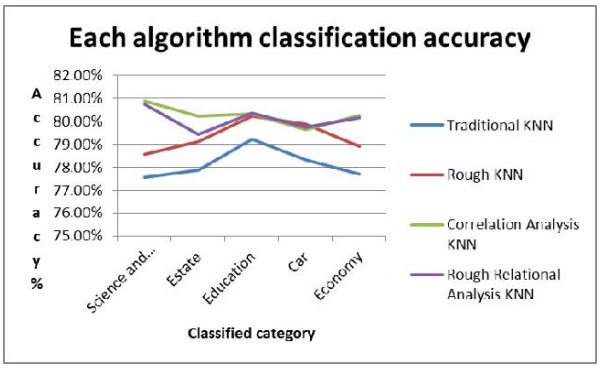


Figure 1. Comparison of the accuracy of each classification algorithm

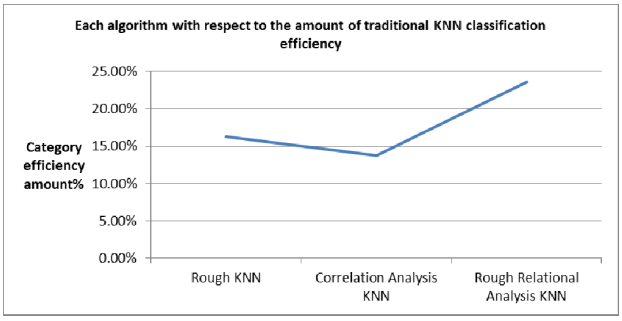


Figure 2. comparison of classification efficiency for four algorithms

From the results in Figure 1, we can find KNN text classification algorithm based on rough sets and association analysis and KNN text classification algorithm based on association analysis have higher accuracy in science and technology, real estate and economics, comparing with two other algorithms. Combined with these three characteristics, these three categories of features are same partly. This can result in an error in classification process for three algorithms. Because improved KNN text classification algorithm carries on association analysis for text vector of each category. We use the average of the sample to be classified text and neighbors of each category will be category determining similarity. It can reduce the incidence of false positives text category.

Data observed in Figure 2, we can see that the classification efficiency and improving the efficiency of KNN text classification algorithm based on rough sets and associated analysis is superior to the rough-based KNN text classification algorithm text classification algorithm based association analysis, and it has the absolute advantage effectively.

V. CONCLUSION

Experimental results show that in contrast to, the improved algorithm KNN text miss a little accuracy and reduces the time complexity, greatly improving the efficiency of text classification, compare with the existing text classification algorithm.

However, due to time constraints and capabilities, a lot of work needs to be improved and deepened. There are several aspects to be improved in future research:

- Because the limit of identified and uncertain area are vague, we can research a more effective reasonable way to determine their boundaries;
- A test set don't contain information it contains, we can't find all frequent item sets for the class of text exactly, this will cause some error for subsequent text classification.

REFERENCES

[1] YingxueSu, Yaowen Fu.Combined non-search feature based KNN algorithm selection algorithm [J]. Computer Engineering, 2007, 33 (18): 217-218, 221.

[2] Chen J,Huang H, TianS,etal.Feature selection for text lassification with naïve Bayes[J].Expert Systems with Applications,2009,36(3):5432-5435.

[3] YupingQin, QingAi, XiukunWang like.Based on Support Vector Machine and class text classification algorithm [J]. Computer Engineering and Design, 2008,29 (2): 408-410.

[4]]Zhang Minling,Zhouzhihua. ML-KNN:Alazy learning approach to multi-label learning [j].Pattern Recognition,2007,40(7):2038-2048.

[5] GuoheFengand,Jingxue Wu .KNN classification algorithm to improve research progress [J] # Library and Information Service, 2012,56 (21): 97-100.

[6] [XiaoyingXu, XiaoyeWang,TaihangDu. Fuzzy ART in the K- Nearest Neighbor Classification Algorithm [J]. Hebei University of Technology, 2004,33 (6).

[7] RongzongSun, DuoqianMiao, ZhihuaWei. Li rough set based on fast KNN text classification algorithm [j]. Computer Engineering .2010.12 (24): 175-177.

[8] [YuanWang, Ye ZhengLiu, YuanchunJiang.Rough KNN text classification algorithm [j].Hefei University of Technology .2014.12 (12): 1514-1517.

[9] XiaodongQian, ZhengouWang.Improved KNN text classification method based on [J].Information Science,2005,23 (4).

[10] HengliangFan, Cheng Weiqing.KNN text categorization method based on relational analysis [j].Computer Technology and Development .2014.6 (6): 72-74.

Project Evaluation of Jilin Rural Power Grid Reformation Based on Rough Set and Support Vector Machine

Du Qiushi
School of Management,
Jilin University
Changchun, China
e-mail: sgccdqs@126.com

Wang guan-nan
Power Economic Research Institute
State Grid Jilin Electric Power
Company Limited
Changchun, China

Cong Li
Jilin Electric Power Company
Limited Information
Communication Company
Changchun, China

***Abstract*—Based on the current situation of rural power grid construction and reformation project in Jilin rural areas, evaluation index system of rural power grid reformation project is established. Aiming at characteristics of a larger number of indicators, rural power grid reformation evaluation model is put forward based on rough set and support vector machine. Through the examples analysis of evaluation data of Jilin rural power grid, rural power grid reformation makes remarkable benefit, proving that the method has higher classification accuracy which is practical and feasible and the model works well.**

Keywords- rural power grid reformation; rough set; SVM(Support Vector Machine)

I. INTRODUCTION

For a long time, extensive management has been using in power grid construction and reformation project in the rural of China, lacking strict and scientific evaluation system of project management. Social and economic benefits have often been ignored when new construction project sets. In this paper, evaluation system of comprehensive analysis is introduced into power grid project, to get scientific and objective conclusion.

To establish a scientific evaluation index system contributes to comprehensive evaluation from society, economy and benefit and effect of a project, especially for construction and reformation project of rural power grid, which has an obvious social public welfare. Practical and innovative index system is needed to come to evaluation and conclusions scientifically and accurately, thus providing an important reference for management decisions.

In this paper, evaluation index system of power grid construction and reformation in Jilin rural areas is built. Evaluation model of rural power grid reformation project is put forward based on rough set attribute reduction and SVM classification, for rough set is used to make up for SVM deficiency in reducing redundant information, and SVM is used to make up for rough set deficiency in generalization ability. On the basis of the principle of rough sets and support vector machine, two types classified study of project evaluation carries on, and empirical research by actual data is conducted, proving that this method has higher classification accuracy.

II. EVALUATION INDEX SYSTEM OF RURAL POWER GRID REFORMATION PROJECT

Building scientific and perfect evaluation index system is an important prerequisite to accurate evaluation of rural power grid reformation project and the basis of comprehensive assessment. Regarding the characteristics of power grid construction and reformation project in Jilin rural areas, through analyzing a variety of factors of rural power grid construction and reformation, index system is built including power grid property, regional economic, social benefit and enterprise finance (Table I).

TABLE I. EVALUATION INDEX SYSTEM

power grid property	line losses rate P_1
	power grid power factor P_2
	power grid security P_3
	user terminal voltage qualified rate P_4
	power grid distribution P_5
	capacity-load ratio P_6
	electric reliability P_7
regional economic	primary industry value P_8
	secondary industry value P_9
	tertiary industry value P_{10}
	proportion of secondary and tertiary industry value P_{11}
	investment climate improvement P_{12}
social benefit	Average alleviate burdens P_{13}
	Peasant satisfaction P_{14}
	Average domestic consumption P_{15}
	government satisfaction P_{16}
	Living environment improvement P_{17}
enterprise finance	internal rate of return P_{18}
	pay back period P_{19}
	Financial Net Present Value P_{20}

III. BASE ON ROUGH SET THEORY AND ATTRIBUTE REDUCTION

A. Knowledge Representation System

Based on rough set, sample set is abstracted as a decision making system, $S=(U,\{V_a\},a)$, where, U is nonempty finite

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.128

set, naming domain of discourse; A is nonempty finite set, naming property set; $\{ V_a \}$ is range of $a \in A$; $a:U \rightarrow V_a$ is injection, which makes any element in domain of discourse U have unique value for a in Va. If A consists of condition attribute set C and decision attribute set D, which satisfy $C \cup D=A$, $C \cap D=$, S is named decision making system[1].

B. Attribute Reduction

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

In selecting various kinds index, not all indexes are important, while some of them are redundancy. Attribute reduction means to delete uncorrelated and unimportance attribute, in the condition of maintaining attribute classification conditions unchanged. As for a c, if $pos_c(D_j)=pos_{c-\{a\}}(D_j)(pos_c(D_j))$, ($D_j$ is positive region of c)，a is redundancy, and $c=c-\{a\}$ is reduction.

C. Decision Rule Extract

$S=(U,\{V_a\},a)$ is a knowledge representation system. $a =$ C D，C is condition attribute set, D is decision attribute set. Knowledge representation system including condition attribute set and decision attribute set is decision tables. X_i and Y_i respectively represent each equivalence class of U/C and U/D. $des(X_i)$ and $des(Y_i)$ respectively represent description of equivalence class X_i and Y_i, that is specific property value of equivalence class X_i and Y_i.Definition of decision rule is as follows: r_{ij}: $des(X_i) \rightarrow des(Y_i)$.Regular certainty factor is as follows: $\mu(X_i, Y_i)=card(Y_i, X_i)/card(X_i)$, where, $0<\mu(X_i, Y_i)\leq 1$; when $\mu(X_i, Y_i)=1$, r_{ij} is certain; when $0<\mu(X_i, Y_i)<1$, r_{ij} is uncertain.

In a decision making system, dependency or correlation exists among each condition attribute in some extent. Reduction can be interpreted that conclusion attribute of decision making system has dependency and correlation upon condition attribute set in the simplest way under a non-losing information premise. Among decision making system, if attribute significance is more, influence of decision partition by attribute is more, and relatively decision attributes is more significant [3].

IV. THEORY AND ALGORITHM OF SUPPORT VECTOR MACHINE

SVM is a common type of feedforward networks, for which the main idea is to create a super-flat surface as a decision hook face, making both sides tend to the outer edge of the septal lines. This induction principle is based on the fact that learning machine error rate on the test data is bounded by training error rate and the sum of terms depending on Vapnik Chervonenkis dimension. In the separable model, the former term of SVM is zero, and the second term is minimized.SVM provides a good generalization ability in the pattern classification, which is the unique property of SVM. The basic idea can be illustrated by the two-dimensional case shown in Fig.1.In Fig.1, triangles and squares represent two classes of samples, which can be divided by the hyperplane H. H1 and H2 are two bounding hyperplane which are the closest samples in the two classes of samples and parallel to the hyperplane H, and the distance between them is called margin. The so-called optimal hyperplane claim not only the hyperplane can separate two classes correctly (training error rate is 0), but also make the margin largest [4], [5].

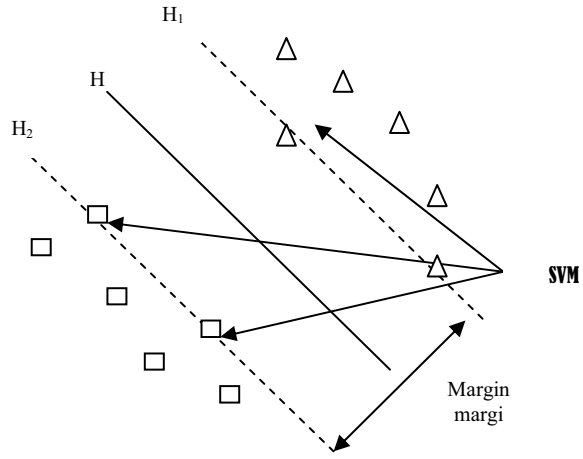


Figure 1. Optimal classification face In linear detachable condition

A. Linear Condition

In linear detachable condition, sample set is sample *set:* (X_i, Y_i), $i=1,2,\ldots,n$, x R^d, and classes can be expressed by y_i $\{+1,-1\}$. Classification linear equation is $w \bullet x_i+b=0$, for which being normalizing, to get y_i $(w \bullet x_i+b)\geq 1, i=1,2,\ldots,n$. When margin is $2/$ w , maximum margin makes w^2 least. Classification face meeting the above conditions and making w^2 least becomes the optimal classification face, and training sample points in H1 and H2 are called support vector. Solving the problem of optimal hyperplane can be expressed as getting the following function minimization:

$$\phi(w)=\frac{1}{2}w^T w \tag{1}$$

Meet constraint condition as follows:

$$y_i\,(w \bullet x_i+b)\geq 1, i=1,2,\ldots,n. \tag{2}$$

This quadratic programming problem can be turned to its antithesis problem, which is seeking the maximized objective function:

$$Q(\alpha) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j x_i^T x_j \qquad (3)$$

Where, Lagrange coefficient $\{a_i\}^n$ should meet constraint condition:

$$\sum_{i=1}^{n} a_i y_i = 0 \text{ 且 } \alpha_i \geq 0 \quad i = 1, 2 \cdots, \text{ n} \qquad (4)$$

Where, $a_i \geq 0$ is introduced to solve the Lagrange coefficient. The problem is to optimize quadratic function under inequality constraints, which exists unique solution. To prove, only a part of Lagrange coefficient in solution is not zero, which corresponds support vector support vector. If a_i^* expresses the optimal Lagrange coefficient, the optimal weight vector can be expressed as:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \qquad (5)$$

The optimal offset can be expressed as:

$$b^* = y_j - \sum_{i=1}^{n} y_j \alpha_i^* x_i^T x_j \qquad (6)$$

The optimal classification function can be expressed as:

$$\begin{aligned} f(x) &= \operatorname{sgn}\left[\left(w^* \bullet x\right) + b^*\right] \\ &= \operatorname{sgn}\left[\sum\ a_i^* y_i \left(x_i \cdot x\right) + b^*\right] \end{aligned} \qquad (7)$$

b* is classification threshold value, which can be get by any support vector, or mid-value of any support vector in classification 2.

B. Nonlinear Condition

As for nonlinear detachable condition, linear problem in a higher-dimensional space is transformed by nonlinear transformation, and optimal classification hyperplane is seeking in transformation space. In the solution process, inner product kernel function K(xi, xj) satisfying Mercer conditions is used properly can realize linear classification after nonlinear transformation. At this point, the question becomes seeking for maximum objective function:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (8)$$

Where, Lagrange coefficient is $\{a_i\}^n{}_{i=1}$. Classification function satisfying constraint condition is as follows:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x_j) + b^*\right\} \qquad (9)$$

which is support vector machine.

Kernel function in this paper is radial basis kernel function:

$$K(x, x_i) = \exp\left[-\frac{\|x - x_i\|^2}{2\sigma^2}\right] \qquad (10)$$

V. EMPIRICAL RESEARCH

A. Instruction of Data Acquisition

The data in this paper come from Jilin Electric Power Company and other related power industry. Method of scoring by specialist is used for qualitative evaluation. Evaluation forms are distributed to relevant experts of Electric Power Company, electric power research institute and electric power design institute, who are asked to make an objective scoring. Each indicator can be any value between 0 to 1.Review levels of power industry credit evaluation index are {very bad, bad, fair, good, very good},and corresponding evaluation scores are {(0-0.2), (0.2-0.4), (0.4-0.6), (0.6-0.8), (0.8-1)}. For quantitative indicators, actual data are used to carry through normalization processing, and then get 20 sets of experimental data.

B. Rough Set Attribute Reduction

Because rough set can only deal with discrete data, equal frequency method is used to discretize data in this paper. Equal frequency method is divide m objects into k segment by parameter k given by users, and each period has m/k objects. Assumption is that the maximum attribute value is $x_{\max}$, the minimum attribute value is $x_{\min}$, and parameter k is given by users. Breaking point set is got by division of k segment of values of all living examples in this attribute averagely, which is arranged from small to large. The number of attribute value between adjacent breakpoint is equal.

In this paper, Johnson's Algorithm is used to carry through attribute reduction for decision data sheet, to get reduction set {line losses rate, user terminal voltage qualified rate, electric reliability, secondary industry value, proportion of secondary and tertiary industry value, average alleviate burdens, peasant satisfaction, average domestic consumption, living environment improvement, internal rate of return, pay back period, financial net present value}.The property after reduction is used as SVM input, which is being trained and tested.

TABLE II. INDEX VALUE AFTER REDUCTION

	After	Before	Before	After
1	0.88	0.68	0.82	0.90
4	0.94	0.82	0.79	0.88
7	0.81	0.83	0.75	0.79
9	0.91	0.83	0.82	0.72
11	0.91	0.82	0.85	0.93
13	0.96	0.46	0.52	0.98
14	0.88	0.52	0.51	0.84
15	0.92	0.77	0.81	0.91
17	0.92	0.90	0.93	0.96
18	0.96	0.93	0.92	0.97
19	0.92	0.91	0.93	0.94
20	0.93	0.88	0.92	0.94

C. *Classification of SVM*

1) Option of training set and test set

Option of training sample.8 groups of training samples are chosen to investigate if sample data can be classified effectively by SVM. Proportions of each group train samples in the total samples are different, and other samples are test samples. Data in each group of training samples and test samples are used to carry through normalization processing.

In this paper, Jilin rural power grid construction and reformation project is divided into two categories: satisfaction and disappointment. Satisfaction means that effect of Jilin rural power grid reformation is well, and rural power grid will make considerable progress and development achieved with the same electricity price between urban and rural areas. Disappointment means that rural power grid have not achieved the expectant results, and the situation of rural power grid did not change significantly. In SVM method, +1 is on behalf of satisfactory for Jilin rural power grid reformation project, and -1 is on behalf of disappointment for Jilin rural power grid reformation project.

2) Classification training of SVM

Svmdark software is used to carry through SVM classification training for each group of training samples. Parameter C and g choose different values depending on each group of training samples. Proportion of training samples which is 30% of the total samples is chosen as an example. 8 groups of training samples are chosen as input for svmdark software and C = 98.158308, g = 0.700168 as model parameters is selected. Radial basis kernel function is used to calculate for each group of training samples.

SVM method is used for classification for corresponding proportion of each group of test samples. The result is as follows (Table ш):

3) Interpretation of result

As can be seen from the above experimental results, the classification accuracy of training and test samples have reached a higher rate. When proportion of training samples which is about 30% of the total samples, the classification accuracy of training reaches maximum. When the proportion of training sample increases, the classification accuracy of test samples shows a downward trend. Classification accuracy of each test samples has reached more than 60%, and average classification accuracy has reached more than 85%. This shows that the effect of SVM for Jilin rural power grid construction and reformation project evaluation is well. Moreover, software's training speed is fast for training samples in the actual operation, and good effect of predicting the unknown samples in the use of small samples shows that the method has a strong practical applications.

TABLE III. CLASSIFICATION ACCURACY OF SVM

Proportion of training samples	Classification accuracy of training samples	Classification accuracy of test samples	Average classification accuracy of all samples
20%	0.8265	0.8478	0.8298
30%	0.8436	0.8797	0.8276
40%	0.9218	0.9432	0.9308
50%	0.9264	0.9102	0.9176
60%	0.9216	0.8545	0.8906
70%	0.9362	0.8306	0.9042
80%	0.9089	0.8869	0.9032

VI. CONCLUSION

In this paper, rough set and SVM model is used to divide effect of Jilin rural power grid construction and reformation project evaluation into two categories: satisfaction and disappointment. Empirical research of expertise data shows that the method has better classification results, which can play a good role in guiding the actual assessment of rural power grid.

The empirical results show that as models based on rough set and SVM apply in Jilin rural power grid construction and reformation project evaluation, because of limited training samples and foundation of nonlinear mapping relation, dimensionality problem is solved. This algorithm has the advantages of simpleness and high accuracy, which can meet the needs of practical application, to provide an effective tool for the rural power grid construction and reformation project evaluation. The evaluation method for comprehensive evaluation of small data samples and high precision requirement has certain significance.

REFERENCES

[1] Ziarko W, "Introduction to the special issue on rough sets and knowledge discovery," Computational Intelligence. Vol.11, 1995.

[2] V.Vapnik. The Nature of Statistical Learning Theory [M].New York:Springer-Verlag.

[3] Li Jian-Ping, Xu Wei-xuan, Liu Jing-li, "The Study of Support Vector Machine in Consumer Credit Assessment," Systems Engineering. 2004.10

[4] Wang Qiang, Shen Yong-ping, Chen Ying-wu,"Model and Algorithm for Multiple Attribute Decision Making Based on Support Vector Machine," Control and Decision. Vol.21, pp.1338-1342,2006..

[5] Zhu Yong-sheng, Zhang You-yun, "The Study on Some Problems of Support Vector Classifier," Computer Engineering and Applications. Vol.13, pp.36-38, 2003.

Acknowledgements:

Project funding: State Grid Jilin Electric Power Company Limited 2014 Technology Project funding, Research on Total Factor Power Energy Efficiency and its influencing factors. Item No. 5223001352CG

The Comprehensive Evaluation Index System for Huadian Transformer Substation Address Selection Based on AHP and SVM

Du Qiushi
School of Management,
Jilin University
Changchun, China
e-mail: sgccdqs@126.com

Miao Qian
Power Economic Research Institute
State Grid Jilin Electric Power
Company Limited
Changchun, China

Cong Li
Jilin Electric Power Company
Limited Information
Communication Company
Changchun, China

***Abstract*—This article builds the system of comprehensive evaluation index system for transformer substation address selection with the analysis of Huadian project for power transmission and transformation, due to the comprehensive evaluation of power transmission and transformation project involving multiple factors, analytic hierarchy process(AHP) is used to determine the weight of each evaluation index to make the combination with subjectivity and objectivity, introduces support vector machine(SVM) into comprehensive evaluation of power transmission and transformation and sets Huadian project as an example for empirical study, the project research is used to improve power supply capability and reliability of central Jilin grid.**

Keywords- support vector machine; analytic hierarchy process; transformer substation ; evaluation;

I. INTRODUCTION

Huadian transformer substation belongs to Jilin region, it needs to give consideration to many facts such as urban planning, environmental protection, military installations, territorial resources, aviation, historical relic and so on, so the project is difficult. According to the overall planning of Jilin west power grid and the principle of transformer substation address selection, it builds the system of comprehensive evaluation index system from four aspects: geology, engineering technology, construction factors and economy based on the analysis of many relevant data. Using AHP and SVM for transformer substation address selection to choose the suited address and verify the accuracy of the model through the empirical research. In the meantime it provides important reference for address selection of Huadian transformer substation.

II. COMPREHENSIVE EVALUATION INDEX SYSTEM

In view of the characteristics of Huadian transformer substation site, it builds the system of comprehensive evaluation index system from several aspects: geology, engineering technology, construction factors and economy based on the analysis of relevant data. The main factors are all evaluation index of the system (Table I).

III. THEORY OF AHP

A. AHP

AHP was put forward by T.L.Saaty the American expert in operational research In the 70s. The basic idea of AHP is to divide complex things into orderly hierarchies and build a standalone hierarchical structure(model tree) to describe system function or characteristic, then give a quantitative representation based on the importance of each hierarchy through a judgment of objective things, it means comparison judgment matrix, using the max eigenvalue and homologous eigenvector determine the weight of every factor about relative important order in each hierarchy that on the premise of through the consistency check; through the analysis of each hierarchy to educe the analysis about the whole problem, it means total sequencing weight. AHP makes the thinking process hierarchical and quantitative and provides quantitative evidence by using mathematical methods, so AHP is a qualitative and quantitative method to give analysis for weight. [1], [5]

TABLE I. COMPREHENSIVE EVALUATION INDEX SYSTEM

primary indexes	geology P_1—P_9	
	engineering technology P_{10}—P_{13}	
	construction factors P_{14}—P_{17}	
	economy P_{18}—P_{20}	
secondary indexes		
geographical location P_1		workload of 500kV, 220kV line export in forward period P_{11}
topography P_2		traffic Transportation P_{12}
geology P_3		ground treatment P_{13}
source of water P_4		environmental conditions P_{14}
line export P_5		execution conditions P_{15}
system conditions P_6		quantity of earthwork P_{16}
flood control and drainage P_7		remove and compensation P_{17}
impact on the communication P_8		Line project investment in current and forward period P_{18}
reserve power supply P_9		comparison of short-term investment than the difference P_{19}
workload of 500kV, 220kV line export in current period P_{10}		comparison of long-term investment than the difference P_{20}

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.129

B. Weight determination

Using AHP to determine weights so that to build the system of comprehensive evaluation index system, even if the indexes are associated or information overlapping, it would not affect the conclusion greatly about the comprehensive evaluation because the total weight has been determined in the same hierarchy.

C. Elementary step of AHP

- Hierarchical structure model. The most important step is to build hierarchical structure model in AHP, it divides the factors into different hierarchy after the thorough analysis, and the hierarchy including top, middle and bottom layer. The top layer is target layer which means the decision maker's goal; the middle layer is criterion layer which to indicate whether objective achieved, the bottom layer is index layer which means the judgment of index.
- To establish the judgment matrix. The relation of the factors between the three layers has been defined after the hierarchy model was set up. We need to estimate relative importance of various factors in each hierarchy. Using numerical representation through bringing in appropriate scale in order to quantify judgments in AHP which was written judgment matrix A.

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots\cdots & a_{1n} \\ a_{21} & a_{22} & \cdots\cdots & a_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ a_{n1} & a_{n2} & \cdots\cdots & a_{nn} \end{bmatrix}$$

- Calculating weight vector. We need to calculate each weight vector in order to extract useful information from judgment matrix so that to reach the cognition about objective things then provide scientific basis for decision making. Setting $A=\left[a_{ij}\right]_{n\times n}$, if $\forall i,j,k=1,2,...,n,\ a_{ik}=a_{ij}a_{jk}$ is right, A is consistency matrix. The factors in A was represented as $a_{ij}=\frac{w_i}{w_j}$. A do not always satisfy consistency conditions, but we can put forward the following method to solve weight vector referring to the character of consistency matrix. Summing every factors in A

$$\bar{w}_i=\sum_{j=1}^{n}a_{ij}\quad i=1,2,...,n \tag{1}$$

Then normalizing, weight vector was

$$w_i=\frac{\sum_{j=1}^{n}a_{ij}}{\sum_{k=1}^{n}\sum_{j=1}^{n}a_{kj}} \tag{2}$$

IV. SVM CLASSIFICATION THEORY

SVM is developed on the basis of statistical learning theory, it's the concrete realization of the VC dimension theory of statistical learning theory and minimum structural risk principle. The distinctive features of this method is to use a few support vectors represent the whole sample set to classify unknown samples. The basic idea can be shown as two dimensional cases in figure 1. Round and square represent two classes of samples, and we can use hyperplane H to partition them. H1 and H2 are the nearest samples from H in two kinds of samples and they are boundary hyperplane which parallel to H. The distance between them is called classification margin. The so-called optimal classification hyperplane is required not only to separate two kinds of samples correctly (training error rate is 0) but also to maximize classification margin. The vector nearest to optimal classification hyperplane is called support vector. According to the characteristics of the data, here introduce two kinds of situations in the following: the linear and nonlinear. [2], [3]

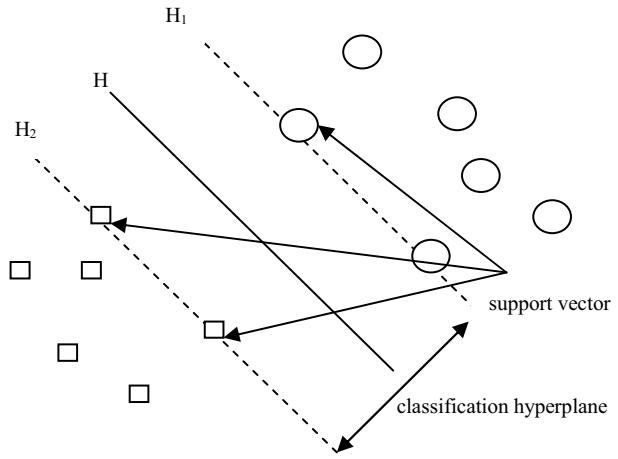


Figure 1. The basic idea of SVM

A. The linear

Setting the input training sample is $x_i\in R^n$, $i=1,2\cdots,\ n$, and the corresponding expectations is $y_i\in\{+1,-1\}$, here + 1 and 1 respectively represent the sign of two kinds of samples. Assume the classification hyperplane exist, the formula is $w*x_i+b=0$, here w is adjustable weight vector, b is the bias of H. In order to let classification hyperplane make classify all the samples

correct we need to make the classification margin $2/\|w\|$ maximum. So in the case of linear separable, the problem solving optimal hyperplane can be expressed as:

Minimize the following function

$$\phi(w)=\frac{1}{2}w^T w \tag{3}$$

And satisfy the following constraints

$$y_i(x_i\cdot w+b)\geq 1 \quad i=1,2\cdots, \quad \text{n} \tag{4}$$

The quadratic programming problem has been transformed into dual problem，it means to seek the max objective function:

$$Q(\alpha)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j y_i y_j x_i^T x_j \tag{5}$$

Its Lagrange coefficient $\{\alpha_i\}_{i=1}^n$ satisfy the constraint conditions

$$\sum_{i=1}^{n}a_i y_i=0 \text{ ， 且 } \alpha_i\geq 0 \quad i=1,2\cdots, \quad \text{n} \tag{6}$$

Here $\alpha_i\geq 0$ is used for solving Lagrange coefficient. This is a quadratic function optimization problem under inequality constraints and has unique solution. Only a part of Lagrange coefficient is not 0 in the solution and the corresponding sample is support vector.

Arrange α_i^* to be the optimal Lagrange coefficient then the optimal weight vector can be expressed as:

$$w^*=\sum_{i=1}^{n}\alpha_i^* y_i x_i \tag{7}$$

The optimal bias is expressed as

$$b^*=y_j-\sum_{i=1}^{n}y_j\alpha_i^* x_i^T x_j \tag{8}$$

Then we can get the optimal classification function:

$$f(x)=\text{sgn}\left\{\left(w^*\cdot x\right)+b^*\right\} \tag{9}$$

In the case of linear inseparable we can add a slack variable in (4), it's $\varepsilon_i\geq 0, i=1,2\cdots n$ and the question is changed:

Minimize the following function

$$\phi(w,\varepsilon)=\frac{1}{2}w^T w+C\sum_{i=1}^{n}\varepsilon_i \tag{10}$$

And satisfy the following constraints

$$y_i(w^T x_i+b)\geq 1-\varepsilon_i, \text{ and } \varepsilon_i\geq 0 \; i=1,2\cdots, \quad \text{n} \tag{11}$$

Here C is a positive parameter.

The solving process is similar to linear separable cases, only constraints are changed into

$$\sum_{i=1}^{n}a_i y_i=0, \text{ 且 } 0\leq\alpha_i\leq C \quad i=1,2\cdots, \quad \text{n} \tag{12}$$

B. The nonlinear

For nonlinear separable problem we can change it into the linear separable problem in certain high dimensional space by using nonlinear transformation and solve the optimal classification hyperplane H in the transform space. In the process of solving, adopt the appropriate kernel dot product $K(x_i,x_j)$ that satisfied the Mercer conditions then we can achieve linear classification after a certain nonlinear transformation. [4],

Now the problem has been changed:

Searching the Lagrange coefficient $\{\alpha_i\}_{i=1}^n$ for max objective function shown as (13)

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{13}$$

And $\{\alpha_i\}_{i=1}^n$ satisfy the (12). The classification function is shown as

$$f(x)=\text{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x_j)+b^*\right\} \tag{14}$$

It is the SVM.

At present we use kernel function as the following three commonly:

- polynomial kernel function

$$K(x,x_i)=[(x\cdot x_i)+1 \quad d=1,2\cdots \tag{15}$$

- RBF kernel function

$$K(x,x_i)=\exp(-\frac{|x-x_i|^2}{\sigma^2}) \tag{16}$$

- Sigmoid kernel function

$$K(x,x_i)=\tanh[v(x\cdot x_i)+c] \tag{17}$$

Here v and c is constant.

The basic idea of SVM: firstly change the input space into a high dimensional space by using nonlinear transformation to make samples linearly separable; then solve optimal classification hyperplane under linearly separable conditions which achieved through defining appropriate kernel dot product.

V. THE EMPIRICAL RESEARCH

A. Data collection

In this paper the data are from the power company, electric power exploration and design institute and other units. We use expert estimation method and provide evaluation list to experts from the field of transformer substation, engineering technology, geology and they give objective estimation. We can value all the indexes between 0 and 1. The reviews level of evaluation index for Huadian project can be shown as {bad, poor, common, good, wonderful} and the corresponding evaluation score is shown as {(0-0.2), (0.2-0.4), （0.4-0.6）, (0.6-0.8), (0.8-1)}. We use AHP to determine the weight through the 20 groups of experimental data.

TABLE II. THE EVALUATION INDEX DATA (UNIT %)

index	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
weihgt	4	5.6	6.1	3.7	5.2	4.6	3.6	4.1	5.8	6.1
index	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}	P_{19}	P_{20}
weihgt	5.3	4.5	4.2	2.9	2.7	5.5	4.1	8.4	6.7	6.9

B. *Attribute reduction of decision tables*

We calculate attribute reduction of decision tables and get the attribute reduction set {topography P_2, geology P_3, line export P_5, system conditions P_6, reserve power supply P_9, workload of 500kV, 220kV line export in current period P_{10}, workload of 500kV, 220kV line export in forward period P_{11}, traffic Transportation P_{12}, quantity of earthwork P_{16}, Line project investment in current and forward period P_{18}, comparison of short-term investment than the difference P_{19}, comparison of long-term investment than the difference P_{20}}, and they are the input data of SVM, then train and test.

TABLE III. INDEX OF ATTRIBUTE REDUCTION

	1#	*2#*	*3#*
P_2	0.980	0.864	0.879
P_3	0.960	0.854	0.869
P_5	0.905	0.955	0.854
P_6	0.955	0.780	0.930
P_9	0.980	0.930	0.955
P_{10}	0.879	0.930	0.930
P_{11}	0.930	0.960	0.920
P_{12}	0.955	0.905	0.829
P_{16}	0.980	0.955	0.889
P_{18}	0.960	0.829	0.869
P_{19}	0.960	0.879	0.960
P_{20}	0.980	0.829	0.920
decision	1	-1	-1

C. *SVM classification*

- The selection of training set and testing set. The system of comprehensive evaluation index system for transformer substation address selection with the analysis of Huadian project is divided into two levels: ideal and imperfect. The two levels represent if the transformer substation meets the requirements of overall design and comprehensive benefits. We use +1 to represent the transformer substation address selection is ideal, and use -1 to represent the transformer substation address selection is imperfect. Setting Huadian project as an example for empirical study and evaluation index data is divided into two parts. We select 16 sets of data as the training sample, the rest of the four groups of data as test samples.
- Classification training of SVM. Using the software called SVMDARK to give classification training to the samples. Parameter C and g has different values when the training sample is difference. Setting 16 sets of data as the training sample and input SVMDARK. Ultimately we choose C=97.879，g＝0.8236.
  Results analysis. Setting 16 sets of data as the training sample and calculate the rest of the four groups of test samples. The result of the comprehensive evaluation index is 1.089, -0.979, -0.963. The result shows that 1# is better than 2#、3#, it means the project of 1# satisfy the requirements of overall design and comprehensive benefits.

VI. CONCLUSIONS

This article uses AHP and SVM to build the system of comprehensive evaluation index system for transformer substation address selection with the analysis of Huadian project, then the project is divided into two levels: ideal and imperfect, it proved that the method has good classification effect through empirical research and it also has a guiding role to select transformer substation address. Using AHP and SVM to find transformer substation address through limited training samples and establish the nonlinear mapping relationship, also solving the problem of the dimension. This algorithm has the advantages of simple, high accuracy, and it's very suitable for promotion. Therefore, if we input the data of other transmission station to the above model, we can get the comprehensive evaluation index for transformer substation address selection. Decision makers can choose the suitable transmission station address according to the results.

REFERENCES

[1] Qin Shou-kang. Comprehensive evaluation principle and application. Beijing: Publishing house of electronics industry, 2003.

[2] Zhu Yong-sheng, Zhang You-yun. The Study on Some Problems of Support Vector Classifier. Computer Engineering and Applications. 2003.13.

[3] Qi Hen-nian. Support Vector Machines and Application Research Overview. Computer Engineering. 2004.5.

[4] Zhang Xue-gong. Introduction to statistical learning theory and suppor vector machines [J]. Acta Automatica Sinica, 2000，26(1), pp.32-44.

[5] Hu Yong-hong, He Si-hui. Comprehensive evaluation method. Beijing: Science Press, 2000,10.

[6] He Xiao-qun. Method and application of modern statistical analysis. China Renmin University Press. 1998.11.

Acknowledgements:

Project funding: State Grid Jilin Electric Power Company Limited 2014 Technology Project funding, Research on Total Factor Power Energy Efficiency and its influencing factors. Item No. 5223001352CG

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

A novel changeable sliding window method for predicting horizontal displacement of dam foundation

Chenyang Jiang, Feng Xu, Xin Lv
College of Computer and Information
Hohai University, HHU
Nanjing, China
e-mail: rdjcy@163.com

Guoyan Xu, Yingchi Mao, Longbao Wang
College of Computer and Information
Hohai University, HHU
Nanjing, China
e-mail: xufeng@hhu.edu.cn

***Abstract*—The prediction of horizontal displacement is significant important to the safety-control in dam. A number of regression analysis-based methods have been proposed. However, most of the methods have unsatisfied performance on prediction accuracy in case of the sample size is unknown. In this paper, a method based on changeable sliding window is proposed to predict horizontal displacement of dam foundation dynamically, which the optimal fitting model and samples can be selected through a way of multiple iterations. The experiment results show that the fair changing pattern can be discovered by the proposed method, and it is also suitable for the prediction in other area.**

Keywords-regression analysis prediction; safety control; changeable sliding window; optimal fitting

I. INTRODUCTION

Dams are an important part of water conservancy and hydropower project, they play a great role in controlling flood, power generation, storing water, ecological environmental protection, etc. In our country, we built numbers of hydropower projects, which have made great achievements. Over the past decades, tens of thousands of built dams such as Xiaowan station, Sanxia station have brought us huge economic benefits.

With the rapidly development of hydropower, dam safety control has brought people's attention gradually. Dams will deform [1] with time [2], under the effects of various factors such as environment, hydraulic power, geomechanics, etc. If the deformation of dams beyond the range of warning, the lives and property of inhabitant in downstream area will be threatened. In history, the dam-breaking events not only bring us a crucial lesson, but also remind us the importance of dam safety monitoring and the warming of extreme events. The horizontal displacement of dam foundation is one of the most important aspects of dam safety monitoring. The main basis for evaluating whether the dam is safe is the displacement of upstream and downstream direction. At present, how to accurately predict the horizontal displacement of dam foundation is the key point of our research.

For the present, in order to analyze the relationship between the horizontal displacement of dam foundation and environmental variables, a great deal of methods such as regression analysis models, timing analysis models and grey models (GM) [3] are proposed. However, most of them have shortages. Regression analysis-based methods have low-performance on prediction accuracy if the sample size is too small [4]. Timing analysis models are lack of ripe research on adaptability and temporal spacing [5]. GM have low prediction accuracy in case of catastrophe [6]. To solve the problems existed in the available methods, a new algorithm named changeable sliding windows is proposed, increasing the accuracy of the prediction.

II. RELATED WORK

Dam safety monitoring system mainly based on boundary conditions (temperature, rainfall, water level, etc.) and structural response (displacement, rotations, pore pressures, etc.) [7]. Deterministic and statistical methods have been applied to predict the behavior of dam several years ago [8]. In the 1950s, multiple linear regression (MLR) models and artificial neural networks (NN) models have been widely used in lots of famous dams. Mata J [9] discussed an important part of artificial neural network (NN) : the parallel processing of the information. Mata J compared MLR with NN models for the characterization of dam behaviour under environment loads. The results of this study reinforce the notion that statistical models are useful for establishing relations between loads and structural responses for the behaviour analysis in the safety control of concrete dams. Stojanovic et al. [10] introduced an adaptive system for dam behavior modeling which is based on MLR models and genetic algorithms (GA), this method need to use the statistical significance of individual regressors to improve the complexity of standard. A method based on neuro fuzzy identification was proposed by Rankovic et al. [11] to predict the horizontal displacement of dam. However, this method did not consider the mechanical properties and any other possible damage directly. Perner et al. [12] analyzed the deformations of arch dam based on hybrid models. This method has the advantage of resulting in good predictions with only a small number of parameters for the MLR-analysis. The existed methods have problems, we combine sliding windows with regression analysis to solve the problems. Sliding windows [13] had been widely used in control-flow. Combined data streams with

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.130

regression analysis prediction, a time series prediction model [14] based on sliding windows was proposed. In this paper, while the amount of data is large, on the basis of previous researches, a new method based on sliding windows is proposed to accurately predict the horizontal displacement of dam foundation.

III. PRELIMINARIES

A. Basic regression theory

Regression analysis is a statistical analysis method, which is widely used to analyze the relationship of two or more variables and establish mathematic model to fit the data.

There are two types of regression analysis: linear regression analysis and non-linear regression analysis. While the sample data is linear, linear regression analysis is suitable. The formula of linear regression is shown as the following:

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \ldots + \beta_i \mathrm{x}_i + \varepsilon \quad (1)$$

$$\mathrm{y} = X\beta + \varepsilon \quad (2)$$

Least square method is an appropriate evaluation criterion which can be applied to estimate β . The fundamental of this criterion is to minimize the root mean square error (RMSE) of sample data, the cost function can be written in the form:

$$\mathrm{J}(\beta) = \frac{1}{2k} \sum_{i=1}^{k} \left(X^{(i)} \beta - y^{(i)} \right)^2 \quad (3)$$

where k is the number of sample. The vector style of the cost function can be shown as the following:

$$\beta = \left(X^T X \right)^{-1} X^T y \quad (4)$$

This system uses quadric polynomial regression model and exponential regression model. Polynomial regression is the promotion of linear regression, the formula is shown as the following:

$$\mathrm{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \quad (5)$$

let $x_1 = x, x_2 = x^2$, quadratic polynomial can be transformed into quadratic linear regression, the value and the method of estimation of β keep fixed.

In the mathematics field, exponential equals to power, a type of computing format of rational power. The general formula of exponential regression is shown as the following:

$$\mathrm{y} = \alpha \beta_1^{x_1} \beta_2^{x_2} (\alpha > 0 \text{and } \alpha \neq 1) \quad (6)$$

where α is exponential regression coefficient, β_1 and β_2 is exponential partial regression coefficient. In this paper, there is only one independent variable x , the formula of exponential regression is shown as the following:

$$\mathrm{y} = \alpha \beta^{\mathrm{x}} \quad (7)$$

B. Asymptotic sampling regression

The sample data of time series data has the property of immediate. In the most of cases, the method has low performance on prediction accuracy if the sample deviates from the target point, specially in short-term prediction.

In order to select better sample, asymptotic sampling considering the time sequence of data. If the sample size is too small, the tendency can't be reflected completely. The sample size should be increased gradually to select an appropriate sample size.

The mathematic description of asymptotic sampling regression is shown as the following: time series data is

$$D = \{ (\mathrm{x}_i, y_i) | i = 1, 2, \ldots, \mathrm{n} \} \quad (8)$$

where x is variable, y is dependent variable, y_m is the $mth(m < n)$ element in set D . A subset which is nearest to element m on time is chosen to predict the response variable y_m .The subset contains k elements:

$$S = \{ (\mathrm{X}_{m-i}, y_{m-i}) | i = k, k-1, \ldots, 1 \} \quad (9)$$

then increase the sample size gradually to select a best sample set

$$S^* = \{ (\mathrm{X}_{m-i}, y_{m-i}) | i = K, K-1, \ldots, 1 \} \quad (10)$$

The optimal estimation can be obtained, and the sample size is the most appropriate.

IV. DYNAMIC DATA FITTING METHOD BASED ON CHANGEABLE SLIDING WINDOWS

A. Algorithm description

In order to keep the safety of dams, data analysis is necessary.Due to the data with higher volume and complexc ity, traditional data analysis algorithm has unsatisfied performance on prediction accuracy. In this paper, a new dynamic data fitting method based on changeable sliding windows is proposed to solve the existed problems. This method uses quadric polynomial regression model and exponential regression model to predict the horizontal displacement of dam foundation under the influence of water level.

The basic of the new method is sliding windows. In the sliding windows mechanism, senders can change the size of window to control flow according to the confirmed information. Regression analysis models have high requirement on data size. Regression analysis models are suitable for predicting the displacement of dam foundation while the data size of dams is large. Changeable sliding windows method makes use of the advantages of two methods to adjusts the sample size. The sample size just means the window, the window can slide with the change of target point.

This algorithm can select sample size dynamically. A static sample size is not applicable to all the points. Dams should be monitored every day, the data size is very large,

so the sample size must be changeable. The sample size too small or too large will lead to under fitting or over fitting. In order to predict accurately, appropriate sample size should be selected for every point. This algorithm can select the size and location of windows every time, then predict the next point automatically. This algorithm runs without manual intervention, it is automated and accurate.

B. Dynamic selection of sample size

Considering the features of quadric polynomial regression model and exponential regression model, the minimum of sample size is 4. As is mentioned above, the three coefficients of quadric polynomial regression is b_0, b_1, b_2, so the minimum of sample size is 3. However, in order to avoid under fitting, the minimum of sample size should be larger than 3. Therefore, when the number of non-empty sample data is larger than 3, this algorithm will have better performance on prediction, so the minimum of windows is 4.

The result of prediction will be more accurate through a way of multiple iterations. RMSE is calculated every time, while RMSE converges to a stable state, RMSE and sample size can be selected at the same time.

C. Dynamic selection of models

This system uses quadric polynomial regression model and exponential regression model to predict the horizontal displacement of dam foundation under the influence of water level. The curve of displacement is similar to the curve of quadric polynomial and exponential, so quadric polynomial regression and exponential regression are more suitable in this system.

After importing data, this system calculates RMSE with two models through a way of multiple iterations. The value of RMSE of two models are different while the sample size is the same. System compares the two values and selects the smaller one, at the same time, the corresponding model is the best model.

D. The determination of evaluation criterion

The smallest value of RMSE is selected as the best result. The formula of RMSE is shown as the following:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_1^n (y_i - \mathrm{y})}{n-1}} \qquad (11)$$

Where n is the number of sample, is the estimated value of present horizontal displacement, is the mean value of the estimated value, n-1 is the degree of freedom of RMSE. Absolute error is the absolute difference of predicted value and measured value, it can also reflect the results just like RMSE. However, in order to avoid over fitting, RMSE is selected as the evaluation criterion. According to law of large numbers, the hypothesis stochastic error term meets the hypothesis of normal distribution of "zero-mean" and "homoscedasticity". Compared with absolute error, RMSE is more suitable for be a evaluation criterion.

E. Convergence

In practice, if the sample deviates from the target point, it will have little influence on prediction, which means the sample size should be small. In order to control the size of sample, this algorithm restricts the times of iterations.

If RMSE is convergence, the iterations will stop. With the increasement of sample size, if RMSE decreases, then keep on iterating; if RMSE does not decrease during 7 times of iterations, RMSE is convergence. Then stop iterating and select the better model.

The times of iterations is controlled on the basis of Bernoulli trial. An iteration can be defined as a Bernoulli trial, each trial is independent of each other. For each trail, the probability of reaching the optimal result is 1/2. If RMSE can't decrease during 7 trials, the probability of getting better result after 7 trials is less than 0.00391(2^(-8)). It is a rare event while the significance level is 0.01, so the result can be selected as the best result.

V. EXPERIMENT

The test platform and parameter of this experiment: (1)CPU is Intel Core i3; (2)memory size is 4G; (3)operation system is Windows7, the software is MATLAB.

Based on the prediction methods introduced in the previous chapters, we choose the data of a dam from January 1,2011 to January 1,2013 to predict.

A. The analysis of RMSE

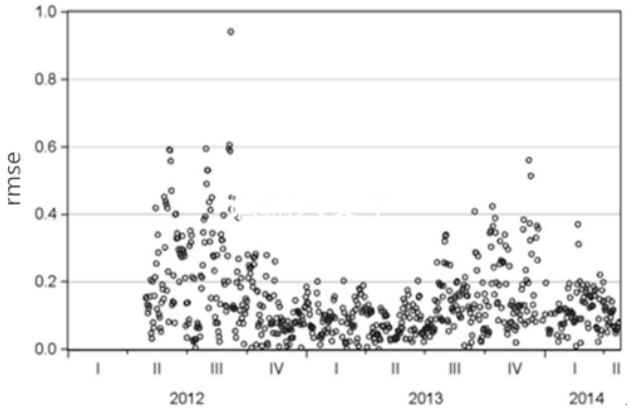


Figure 1: The dot chart of RMSE

The dot chart of RMSE (Figure 1) shows: the three threshold criterions are 0.2、0.4、0.6. As is shown above, about 95 percent of RMSE are less than 0.4mm. The horizontal displacement of dam foundation in this experiment is about 80mm, 0.4mm is only about five-thousandth of 80mm, so 0.4mm is selected as the criterion. When predicting a point, if the value of RMSE is larger than 0.4mm, the optimal model we select is suspect, the point here should be monitored.

TABLE I. The cumulative distribution of RMSE

RMSE	Quantity	Cumulative percentage (without null values)	Cumulative percentage (with null values)
<0.6	679	98.69%	93.40%
<0.5	670	97.38%	92.16%
<0.4	654	95.06%	89.96%
<0.2	537	78.05%	73.87%
null	39	----	5.36%

B. The analysis of error

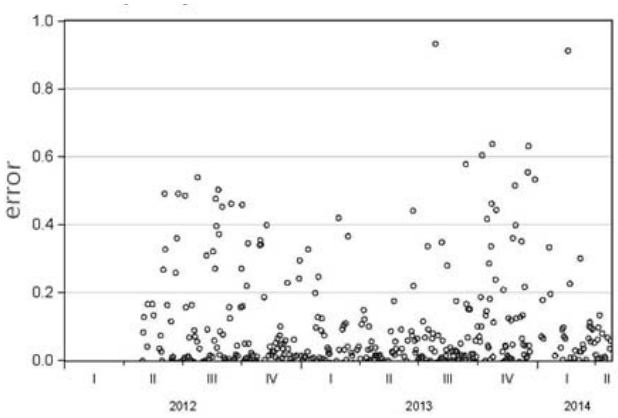


Figure 2: The error of c4-a22-pl-01

In order to explore the prediction ability of this system, we test the next point and get the figure of error(Figure 2)and statistical table(Table II), about 91.7% of error are in the range of ±0.4mm, the result is accurate.

TABLE II. The cumulative distribution of error

Error	Quantity	Cumulative percentage (without null values)	Cumulative percentage (with null values)
<0.6	659	95.92%	90.77%
<0.4	630	91.70%	86.78%
<0.2	561	81.64%	77.27%
<0.01	414	60.26%	56.95%

The experiment shows that the algorithm has satisfied performance on prediction accuracy.

VI. CONCLUSION

It is a challenge to predict the horizontal displacement of dam foundation, people should pay more attention to this problem. Due to the data size of dams is big, a new method based on changeable sliding windows is proposed in this paper to monitor the horizontal displacement of dam foundation. This method uses quadric polynomial model and exponential model to select the optimal fitting model and optimal sample size dynamically. The results of experiment show that this method has high accuracy. This system is also suitable for the prediction of many kinds of data.

This arithmetic still need to be improved because it has low efficiency. When the data size is too big, this algorithm should spend a lot of time to finish the work. In the future research, we will improve computing speed and precision.

ACKNOWLEDGEMENT

This research is partially supported by “National Natural Science Foundation of China” (Grant No. 61272543); “National Key Technology Research and Development Program of the Ministry of Science and Technology of China” (Grant No. 2013BAB06B04); “Key Technology Project of China Huaneng Group” (Grant No. HNKJ13-H17-04); “Natural Science Foundation of Jiangsu Province” (Grant No. BK20130852); “Jiangsu Planned Projects for Postdoctoral Research Funds” (Grant No. 1401001C).

REFERENCES

[1] Bayrak T. Modelling the relationship between water level and vertical displacements on the Yamula Dam, Turkey[J]. Natural Hazards and Earth System Science, 2007, 7(2): 289-297.

[2] Pytharouli S I, Stiros S C. Ladon dam (Greece) deformation and reservoir level fluctuations: evidence for a causative relationship from the spectral analysis of a geodetic monitoring record[J]. Engineering Structures, 2005, 27(3): 361-370.

[3] Li Yang. Applied research of wavelet theory in dam deformation monitoring data analysis[D]. Xi'anUniversity of Technology, 2010.

[4] Xisheng Hua , Teng Huang.Precision engineering measurement technology and application[M].Nanjing:Hohai University Press , 2002:240-245.

[5] Hong Mei. A study of time series analysis method based on robust estimation for deformation monitoring.[D].Nanjing: Hohai University , 2005.

[6] YuhuaZheng.A study of safety monitoring system for a great bridge.[D].Nanjing: Hohai University, 2003.

[7] De Sortis A, Paoliani P. Statistical analysis and structural identification in concrete dam monitoring[J]. Engineering Structures, 2007, 29(1): 110-120.

[8] Ranković V, Grujović N, Divac D, et al. Modelling of dam behaviour based on neuro-fuzzy identification[J]. Engineering Structures, 2012, 35: 107-113.

[9] Mata J. Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models[J]. Engineering Structures, 2011, 33(3): 903-910.

[10] Stojanovic B, Milivojevic M, Ivanovic M, et al. Adaptive system for dam behavior modeling based on linear regression and genetic algorithms[J]. Advances in Engineering Software, 2013, 65: 182-190.

[11] Rankovic V, Grujovic N, Divac D, Milivojevic N, Novakovic A. Modeling of dam behavior based on neuro fuzzy identification[J]. EngStruct 2012;35:107−13.

[12] Perner F, Obernhuber P. Analysis of arch dam deformations[J]. Frontiers of Architecture and Civil Engineering in China, 2010, 4(1): 102-108.

[13] Gibbons P B, Tirthapura S. Distributed streams algorithms for sliding windows[C]//Proceedings of the fourteenth annual ACM symposium on Parallel algorithms and architectures. ACM, 2002: 63-72.

[14] Chaosheng Yan, Chengjiang Zhang, Ying Ma. Study on regression analysis and prediction of time-series data streams using sliding windows[J].Journal of Natural Science of Heilongjiang University, 2007, 23(6): 863-867.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

An Identification Method of News Scientific Intelligence Based on TF-IDF

Lu Pan, Haibo Tang
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, Jiangsu Province, China
pl974473296@163.com

Lei Zhou, Liuyang Wang, Quanyin Zhu*
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, Jiangsu Province, China
hyitzqy@126.com

***Abstract*— With the development of Internet, the amount of Information has been rapidly growing which is spread widely. In order to improve the value and accuracy of science information that is pushed in this paper, an intelligence dichotomous method for science information categorization to identify science information from massive Web news is presents. During the experiment, 85.3% recognition rate of the recognition non-tech news are realized and 82.9% accuracy rate, the results show that the method can effectively identify Web science information news and reduce the amount of independent news.**

Keywords- Scientific intelligence; text categorization; TF-IDF

I. INTRODUCTION

The rapid growth and flow of information has been brought by Internet. Nowadays, the world has entered into a information explosion age, The amount of information stored by computer grows in exponential mode. Science information has important influence on the survival and development of an enterprise, how to find accurate science information from the flood of information has become an key factor for enterprise to adapt to Times change and keep with the Times. Today, Internet has greatly enriched Web information that has different data format and has much redundancy, the text automatic categorization extraction system [2] is an important approach to obtaining science information in need. Text categorization is a key technology for science information categorization extraction, which is applied in areas such as information filtering [10], search engine, discrimination of words, spam filtering, digital library, information retrieval, full text search. The study of text categorization [23] will improve the accuracy of science information categorization, which makes contributions for enterprise to get information timely and efficiently. It is also the best practice for the enterprise's survival and development. These papers design a two classification method for intelligence dichotomous [6] model that can improve the accuracy.

A. Text Categorization

The core of text categorization is to build a function from a single text to category. After years of research, according to the initializing classification number tag learning models [17] can be divided into supervised learning unsupervised, semi-supervised learning, enhance learning and learning.

1) Classification method based on statistics

The basic idea: study on the base of large amounts of text [3], structure characteristic vector and classification mark relationship model, divide texts on the model.

Classification method based on statistical has a solid theoretical basis [24], which is very easy and effective. This method is currently the most widely used text classification [1] method. But it doesn't take the structure of text linguistics into consideration. Instead, the text is represented by a feature vector [12], based on the assumption of independence between words. Text classification based on statistical method [4] is a hot issue, some representative algorithm such as Naive Bayes [5], Class center vector , K-nearest neighbor (K-NN) [18], Maximum entropy modeling [7], Least squares fitting, Support vector machine [15], Artificial neural network [19] and so on.

2) Rule based classification method

Basic ideas: Each classification has its own rules. First, build rule templates under the classification, then count the number of classification rules appearance, according to this information to determine the category of text [22]. Now, many commonly used rule based on classification method such as decision tree, rough set [13] and association rules.

B. TF-IDF

Vector Space Model VSM indicates text by Bag of words model, this Model a Representation that is currently widely used in text classification [16], The basic idea is to change text vector into a vector in a multidimensional space [11], Each dimensional of vectors is a future items and item weights [14]. This method usually takes the following methods to represent the similarity between text: 1) Cosine distance; 2) Euclidean distance. In VSM model, TF-IDF (Term Frequency-Inverse Document Frequency) It is a method of automatically extracting text keywords [20], The method is "word weight" representation, it is the product of the text's local coefficients and global coefficients [12]

$$tf_{ij} * \log(\frac{N}{n_j}) \quad (1)$$

Equation (1) it is the TF-IDF function, TF refers to the importance of the terms for an article, namely Local coefficients; IDF refers to the importance of the lexical items to the training set, It can distinguish between different categories, namely global Parameters. In order to make the boundaries between different types of text clearer, TF-IDF takes the importance of feature words to the single document

*Corresponding author. hyitzqy@126.com

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.131

and the ability to distinguish between word document categories into consideration at the same time. Making the selected feature words [9] can carry more category information.

C. Cosine similarity

Cosine distance, known as the cosine similarity, is a way to measure differences between two individuals by cosine of the angle between two vectors in space.

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} * \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{1}$$

II. EXPERIMENTAL PREPARATIONS

A. Data Preparations

The data of news is used in the experiments, all from these website (http://www.most.gov.cn/, http://www.teda.gov.cn/, http://www.stcsm.gov.cn/, http://www. nsfc.gov.cn/, http://www.hsckjj.gov.cn/, http://www.sjzkj.gov.cn/, http://www.cas.cn/, http://www.gapp.gov.cn/, http://www.sastind.gov.cn/, http://www.jstd.gov.cn/, http://www. kejixun.com/, http://tech.ifeng.com/, http://tech.huanqiu. com/), select the time from January 2012 to June 2015. Every news has three dimensions, include title, keys, content.

B. Experimental Environment

The machine configuration is Windows 7/8, RAM is 4G.

The tool of participle use IKAnalyzer (http: //www.oschina.net/p/ikanalyzer), it is base Lucene.

The news is collect from Web use PyQuery from Python 2.7.

C. Model Training

Definition 1. Set the data set of news collect from Web is given by Equation (3), D_{x_i} has three dimension, include title, keys, content. $D_i = \{title, keys, content\}$.

$$News = \{D_{x_1}, D_{x_2}, ..., D_{x_n}\} \tag{3}$$

Definition 2. Set the artificial classification of news is given by Equation (4), $j + z = x_n$.

$$NC = \{\{D_{n,1}, D_{n,2}, ..., D_{n,j}\}, \{D_{y,1}, D_{y,2}, ..., D_{y,z}\}\} \tag{4}$$

Definition 3. Set the result of text pretreatment is given by Equation (5).

$$DW = \{\{D_{n,1}, D_{n,2}, ..., D_{n,j}\}, \{D_{y,1}, D_{y,2}, ..., D_{y,z}\}\} \tag{5}$$

Each $D_i\left(i \in (j+z)\right)$ is given by Equation (6).

$$D_i = \{W_{p_1}, W_{p_2}, ..., W_{p_n}\} \tag{6}$$

Definition 4. Statistics of word frequency for Equation (7), and the result is given by Equation (8).

$$DF_i = \{W_{x_1} = F_{x_1}, W_{x_2} = F_{x_2}, ..., W_{x_n} = F_{x_n}\} \tag{7}$$

Definition 5. The function of compute local parameter is given by Equation (8).

$$SL_{j,i} = \log(\frac{W_{j,i}}{\sum_{k=1}^{ks} W_{k,i}} + 1) \tag{8}$$

And the result of compute local parameter is given by Equation (9).

$$SL_i = \{W_{x_1} = L_{x_1}, W_{x_2} = L_{x_2}, ..., W_{x_n} = L_{x_n}\} \tag{9}$$

Definition 6. The function of compute overall parameter is given by Equation (10). N is the number of news, F_i is the number of doc for word appear.

$$SG_i = \frac{\log\left(\frac{N+0.5}{F_i}\right)}{\log(N+1)} \tag{10}$$

And the result of compute overall parameter is given by Equation (11).

$$SG = \{W_{t_1} = L_{t_1}, W_{t_2} = L_{t_2}, ..., W_{t_n} = L_{t_m}\} \tag{11}$$

Definition 7. The function of compute weight is given by Equation (12).

$$SLG_j(Wi) = SL_j(W_i) \times SG(W_i) \tag{12}$$

And the result of compute weight is given by Equation (13).

$$SLG_i = \{W_{x_1} = Q_{x_1}, W_{x_2} = Q_{x_2}, ..., W_{x_n} = Q_{x_n}\} \tag{13}$$

The step for model training as follow:

Step 1. Obtain news from Web as data set $News = \{D_{x_1}, D_{x_2}, ..., D_{x_n}\}$. Every news has three dimensions, first dimension is title, the second dimension is keys and the last is content;

Step 2. Evaluation the data set of news through human, the data set of news is divide two categories, tech news and non-tech news $NC = \{\{D_{n,1}, D_{n,2}, ..., D_{n,x}\}, \{D_{y,1}, D_{y,2}, ..., D_{y,z}\}\}$;

Step 3. Text pretreatment for NC_i, the process includes text participle, stop-words process, speech tagging, recognition the name of human and place. Get result is $DW = \{\{D_{n,1}, D_{n,2}, ..., D_{n,j}\}, \{D_{y,1}, D_{y,2}, ..., D_{y,z}\}\}$, and each $D_i\left(i \in (j+z)\right)$ is $D_i = \{W_{p_1}, W_{p_2}, ..., W_{p_n}\}$;

Step 4. The result of statistics of word frequency is $DF_i = \{W_{x_1} = F_{x_1}, W_{x_2} = F_{x_2}, ..., W_{x_n} = F_{x_n}\}$;

Step 5. According to the Equation (8) compute the local parameter of Equation (9), get result is $SL_i = \{W_{x_1} = L_{x_1}, W_{x_2} = L_{x_2}, ..., W_{x_n} = L_{x_n}\}$;

Step 6. According to the Equation (10) compute the local parameter of Equation (11), get result is $SG = \{W_{t_1} = L_{t_1}, W_{t_2} = L_{t_2}, ..., W_{t_n} = L_{t_m}\}$;

Step 7. According to the Equation (12) compute the weight of Equation (13), get result is $SLG_i = \{W_{x_1} = Q_{x_1}, W_{x_2} = Q_{x_2}, ..., W_{x_n} = Q_{x_n}\}$;

Step 8. Save model parameter, include DF, SL, SG, SLG.

D. New Text process

Definition 8. The function of compute similarity of two news text, the vector of news text is weight. The function is given by Equation (13).

$$S = Sim(V_1, V_2) \tag{13}$$

V_1, V_2 is two Vector of news text, and principle is cosine similarity.

The step for model training as follow:

Step 1. Obtain a news from Web is $D_s = \{title, keys, content\}$.

Step 2. Text pretreatment for NC_i, the process includes text participle, stop-words process, speech tagging, recognition the name of human and place. Get result is $D_s = \{W_{p_1}, W_{p_2}, ..., W_{p_n}\}$;

Step 3. The result of D_s statistics of word frequency is $SDF_i = \{W_{x_1} = F_{x_1}, W_{x_2} = F_{x_2}, ..., W_{x_n} = F_{x_n}\}$;

Step 4. According to the Equation (8) compute the local parameter of D_s, get result is $SSL = \{W_{x_1} = L_{x_1}, W_{x_2} = L_{x_2}, ..., W_{x_n} = L_{x_n}\}$;

Step 5. According to the Equation (12) compute the weight of D_s, the overall parameter from model parameter, get result is $SSLG = \{W_{x_1} = Q_{x_1}, W_{x_2} = Q_{x_2}, ..., W_{x_n} = Q_{x_n}\}$;

Step 6. According to Equation (13), compute similarity of D_s's vector with the vector of different news category. The result is S_N and S_Y.

Step 6. If $S_Y > S_N$, add marked of D_s to tech news, otherwise add marked of non-tech news.

III. EXAMPERMENTAL RESULTS

TABLE I. DESCRIPTION OF THE EXPERIMENTAL DATA SET

Dataset name	***Dataset Description***		
	Dimension	***Number of samples***	***Category name***
ScienceData	3	4053	Tech news
NonScienceData	3	3533	Non-tech news
NonClass	3	462	Test data

From Table 1 is Description of the experimental data set. Which is consist of tech news and non-tech ,dimension represents a text consists of two parts: title， keys and content；The number of samples represents the number of Chinese texts.

TABLE II. THE EXPERIMENTAL RESULTS DESCRIBE

Dataset name	***The experimental results describe***		
	Total	***Tech news number***	***Non-tech news number***
Non-tech news	355	6	349
Tech news	107	35	72

Table 2 is the experimental result describe. The result of this experiment is shown in figure 4.3, 41 of 462 samples News is science news, 6 of science news are identified as non-science News, 85.3% recognition rate;

421 of samples News is non-science news, 349 of science news are identified as non-science News, 82.9% recognition rate;

IV. EXPERIMENTS ANALYSIS

Sample set is consisting of 462 articles of news, 41 articles is science-news and 421 articles are non-science articles. This is in line with Internet where much non-tech news exists. After manual testing, the recognition rate of multi-level classification model is more than 80%. Although the recognition rate is not more than 99%, it can provide a reference for news Recommended; meanwhile, multi-level classification model will improve classification efficiency. Expanding scientific and technological vocabulary library can improve the sensitivity to the science and technology news.

V. CONCLUSIONS

Our research work of identification method of news scientific intelligence comes from our related work. In view of the Web having a large number of tech news information. The design has realized the recognition method of science and technology news. Establish science and technology news recognition model by supervised learning way in the Experiment. Then identify new news on this base, experimental scientific intelligence news recognition rate has reached 82.9%, and non-tech news accounts for more than 98.3. The experiment achieved good effect and improves the value of the news collected from Web and the accuracy of science news classification.

APPENDIX

TABLE III. ALL NEWS WEBSITE

People's Republic of China Ministry of Science and Technology	http:// www.most.gov.cn
Tianjin Science and Technology Bureau	http:// www.teda.gov.cn
Shanghai Science and Technology Bureau	http:// www.stcsm.gov.cn
Baoding Science and Technology Bureau	http:// www.bdskjj.gov.cn
The National Natural Science Foundation of China	http:// www.nsfc.gov.cn
Hengshui Science and Technology Bureau	http:// www.hskjj.gov.cn
Shijiazhuang Science and Technology Bureau	http:// www.sjzkj.gov.cn
Chinese Academy of Science	http:// www.cas.cn
China SARFT	http:// www.gapp.gov.cn
Chinese Defense Industry Bureau	http:// www.sastind.gov.cn
Jiangsu Science and Technology Department	http:// www.jstd.gov.cn
Technology News	http:// www.kejixun.com
Phoenix	http:// tech.ifeng.com
Global Technology	http:// tech.huanqiu.com

ACKNOWLEDGMENT

This work is supported by the National Sparking Plan Project of China (2011GA690190), the fund of Huaian

Industry Science and Technology. China (HAG2014023, HAG2014028).

REFERENCES

[1] Zheng-Tao Yu, Lu Han,et al."Study on the construction of domain text classification model with the help of domain knowledge". International Conference on Machine Learning. 2008: 2612 - 2617.

[2] Zhiqing Shao, Hua Nan,et al. "Content-Oriented Automatic Text Categorization with the Cognitive Situation Models". International Symposium on Computer Science and Computational Technology. Kunming: IEEE Computer Society. 2008: 512 - 516.

[3] Jian Yang,Mei Sun,et al."Study on Massive Text Classification Mining Grid System". International Symposium on Information Engineering and Electronic Commerce. 2010: 1-6.

[4] Zhou Faguo, Zhang Fan. "Research on Short Text Classification Algorithm Based on Statistics and Rules". International Symposium on Electronic Commerce and Security (ISECS). Kunming: IEEE Computer Society. 2010: 3 - 7.

[5] Gong Zheng, Yu Tian; "Chinese web text classification system model based on Naive Bayes". International Conference on E-Product E-Service and E-Entertainment (ICEEE). 2010: 978 - 971.

[6] Xinghua Fan, Hongge Hu. "A New Model for Chinese Short-text Classification Considering Feature Extension". A 2010 International Conference on rtificial Intelligence and Computational Intelligence. Kunming: IEEE Computer Society. 2010: 7 - 11.

[7] Ahmadizar, F.,Hemmati, M.."Two-stage text feature selection method using fuzzy entropy measure and an t colony optimization". Iranian Conference on Electrical Engineering. 2012: 695 - 700.

[8] Chupin Chao, Wenbao Jiang."Study on the Subjective and Objective Text Classification and Pretreatment of Chinese Network Text". IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics. Daejeon:Institute of Electrical and Electronics Engineers Inc. 2012: 25 - 29.

[9] Alghamdi, H.S., Tang, H.L.."Hybrid ACO and TOFA feature selection approach for text classification". IEEE Congress on Computational Intelligence. 2012: 1 - 6

[10] Devasena, C.L., Hemalatha, M.. "Automatic Text categorization and summarization using rule reduction". International Conference on Advances in Engineering, Science and Management (ICAESM). 2012: 594 - 598.

[11] Yan Xu.; "A Comparative Study on Feature Selection in Unbalance Text Classification ; Yan Xu". International Symposium on Information Science and Engineering (ISISE). 2012: 44 - 47.

[12] Yan Li, Chungang Chen. "Research on the feature selection techniques used in text classification". International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). 2012: 725 - 729.

[13] Libiao Zhang; Yuefeng Li,et al."Rough Set Based Approach to Text Classification". International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT). Kunming: IEEE Computer Society. 2013: 245 - 252.

[14] Ding Xiaoming; Tang Yan."Improved mutual information method for text feature selection". International Conference on Computer Science & Education (ICCSE). 2013 ; 163 - 166.

[15] Abdul-Rahman, S., Mutalib, S,et al. "Exploring Feature Selection and Support Vector Machine in Text Categorization". IEEE 16th International Conference on Computational Science and Engineering (CSE). Kunming: IEEE Computer Society. 2013: 1101 - 1104.

[16] Mendoza, M., Leon, E.,et al." Clustering of web search results based on an Iterative Fuzzy C-means Algorithm and Bayesian Information Criterion" .IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS). 2013: 507 - 512.

[17] Jiang Xiao-Yu, Jin Shui."An Improved Mutual Information-Based Feature Selection Algorithm for Text Classification". , International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). Kunming: IEEE Computer Society. 2013: 126 - 129.

[18] Jaganathan, P.; Jaiganesh, S. "An improved K-means algorithm combined with Particle Swarm Optimization approach for efficient web document clustering". 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE). 2013: 772 - 776.

[19] Rongze Xia; Yan Jia,et al. "A BP neural network text categorization method optimized by an improved genetic algorithm". 2013 Ninth International Conference on Natural Computation (ICNC). 2013: 257 - 261.

[20] Kewei Shen; Xian Chen."A blended feature selection method in text classification", International Conference on Cyberspace Technology (CCT 2013). 2013: 573 - 576.

[21] Xiaofei Zhou, Li Guo. "Latent Factor SVM for Text Categorization". IEEE International Conference on Data Mining Workshop (ICDMW). 2014 ; Pages: 105 - 110.

[22] Zhe Gao; Yajing Xu,et al. "Improved information gain-based feature selection for text categorization". International Conference on Wireless Communications,Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE). 2014: 1 - 5.

[23] Dai Li, Yi L."Automatic text categorization using a system of high-precision and high-recall models". Dai Li; Murphey, Y.L. ; 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). 2014: 373 - 380.

[24] Nouri, V.; Akbarzadeh-T,et al. "A hybrid type-2 fuzzy clustering technique for input data preprocessing of classification algorithms". 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Kunming: IEEE Computer Society. 2014: 1131 - 1138.

[25] Nouri, V.; Akbarzadeh-T."A hybrid type-2 fuzzy clustering technique for input data preprocessing of classification algorithms": 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2014: 1131 - 1138.

Estimation of Clusters Number and Initial Centers of K-means Algorithm Using Watershed Method

Xiaolong Wang, Yiping Jiao, Shumin Fei
School of Automation
Southeast University
Nanjing, Jiangsu, China
wangxlseu@163.com, jiaoyiping@163.com, smfei@seu.edu.cn

***Abstract*—In K-means clustering algorithm, the selection of cluster number k and initial K-means center has certain influence on the result. It would generate very different aggregation result when confronting with some certain types of data set. This paper aims at proposing an estimation method to evaluate the initial parameters for K-means algorithm. The estimation is executed through data analysis, which contains two main steps: the data would be transformed into data dimensional density first, and then, watershed method would be applied to divide the data space into multiple regions. Each regional center is selected as an initial K-means center, and the number of region is set as cluster number. This estimation method takes advantage of image segmentation ideology and the case study in this paper showed its favorable performance.**

Keywords- K-means algorithm; watershed method; clusters number; initial K-means center

I. INTRODUCTION

Clustering aggregation is a significant problem in pattern recognition, and it's a process to distinguish different objects and to make classification according to certain requirements or rules. The sorting criterion is primary based on the data's similarity of an object in a same category, but there exist less priori knowledge about the data. Therefore, clustering aggregation is a kind of non-supervision classification, which applies mathematical method to study the data [1]. Within these methods, K-means algorithm is quite typical and played a momentous role in both scientific experiments and industrial production. This algorithm is based on the Euler distance of sample points, employing the distance value as evaluation index of comparability. K-means algorithm considers that a category is constituted by close-in-distance object, and the classification target is to obtain compact and independent sorts.

Although K-means clustering shows its efficiency in many traditional applications, its defect appears obviously when the data set become much more complicated. And adopting K-means algorithm to do cluster analysis on these data sets directly is less efficient [2]. This problem is mainly caused by empirical selection of clusters number and random initial K-means center. Some researchers optimized the model of the cloud computing platform to raise the computing speed, and also some proposed initial parametric optimization methods for K-means. Literature [3] selected K-means initial center by analyzing data dimensional density, and literature [4] proposes a new K-means algorithm based on iterative density, which achieved a good performance.

This paper presents a new method to estimate both clusters number and initial K-means center. The method would take advantage of data dimensional density analysis and utilize watershed method to acquire clusters number k and dividing data into multiple regions. The watershed method is mainly used in image extraction, and in this paper, such technique would be used for partitioning data. The number of divided regions could be chosen as k, and its center coordinates as K-means initial centers.

II. THE ESTIMATION METHOD

In this section, the computing process of K-means algorithm is a briefly explained, and then stated the technique transforming data set to density distribution. Eventually, watershed is applied to divide the data region into segmentations, and get the estimation.

A. K-means Algorithm

K-means is a clustering algorithm based on optimizing criterion function. If a data sample is presented as aggregate $X=\{x_1, x_2,\ldots, x_n\}$, x_i is a d-dimensional vector, and suppose the number of clusters is k, the initial K-means center is $C_i(0)$.The similarity measurement adopts Euclidean distance, as for α and β :

$$D = \|\alpha - \beta\| = \sqrt{(\alpha-\beta)^T(\alpha-\beta)} \tag{1}$$

The clustering criterion adopts sum of squared error:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - C_i\|^2 \tag{2}$$

The iterative procedure of K-means algorithm can be summarized as [5]:

- *Step 1.* Parameter initialization: cluster number k is set, and the initial K-means center $C_i(0)$ is set as a random data point. (where j=1,2,…, k)
- *Step 2.* Iterative revision: allocate each data x_i from data set $X=\{x_1, x_2,\ldots, x_n\}$ to a class $C_p(l)$ when

$$\|x_i - C_p(l)\| < \|x_i - C_q(l)\| \tag{3}$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.132

where l stands for iterations. p,q=1,2,…,n, $p \neq q$, i=1,2,…,n.

- *Step 3.* Renew clusters center: the new K-means center at l+1 times calculation is

$$C_j(l+1) = \frac{1}{N_j} \sum_{x_i \in C_j(l)} x_i \tag{4}$$

where N_j is the number of data in cluster j.

- *Step 4.* Termination judgment: if $C_i(l+1) = C_i(l)$ or $|C_i(l+1) - C_i(l)| < \varepsilon$ (ε is allowed minor error), stop calculation and output the result. Otherwise, turn to Step 2.

B. Data relative dimensional density calculation

For the sake of getting a more suitable cluster number k and initial K-means center, the data should be preprocessed. In general, a cluster contains a group of data close in distance or dense in a region. To get such data dimensional density could take advantage of two main means: 1) count the number of data points in a fixed region or 2) apply kernel density estimation [6].

In this paper, the actual density distribution of a data set is not needed in subsequent processing. The data density calculation only needs to provide a relative distribution of data density. A kind of kernel density estimation formula can be simplified in this paper as:

$$d_i = \sum_{j=1}^{n} \exp(-\gamma \left\| x_i - x_j \right\|) \tag{5}$$

where d_i is the intensive degree of data point x_i, and γ is a positive constant. On account of probably uneven distribution, the parameter γ selection could be hard. This paper provides another density estimation method as:

$$d_i = 1 / \sum_{x_j \in S_i} \left\| x_i - x_j \right\| \tag{6}$$

Where S_i stands for a fixed region around data point x_i .The size of this region could set as m nearest other data points ($m=\delta n$).

C. Watershed Method

The data dimensional density can be draw as a grayscale image, which could be cut by watershed method. This method is briefly described as follows:

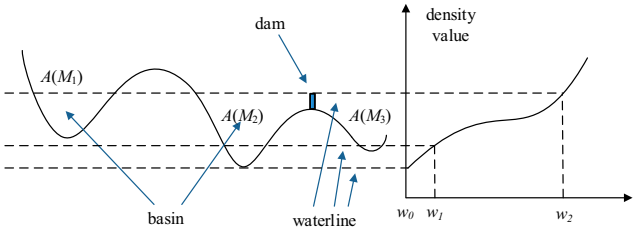


Fig. 1. Principle of watershed method

As is shown in Fig.1, the density value of the data set is sort form small to large, and the data space is regard as a topographic map. When waterline raised from w_0 to w_1 , the area $A(M_2)$ start to impound, and become a basin. When two basin be about to get connected, a dam would be built as an area boundary.

Direct application of watershed method may cause over segmentation, and invalidate the algorithm. This paper utilizes mark to restrict the rise of basin number. Firstly, the 'topographic map' is binarized and disposed by range conversion. The processed 'map' is divided by watershed method and provides a group of parting line I_{em} , which is called extern mark. Then, local minimum value of the 'map' is set as inner mark I_{im} . Taking advantage of the two mark, the data region can be divided [7].

The number of region is set as cluster number k, and each regional center is selected as an initial K-means center.

III. ALGORITHM IMPLEMENTATION

The estimation of initial parameters of K-means algorithm includes two steps as described in prior section. The whole process of the method is summarized as Fig.2.

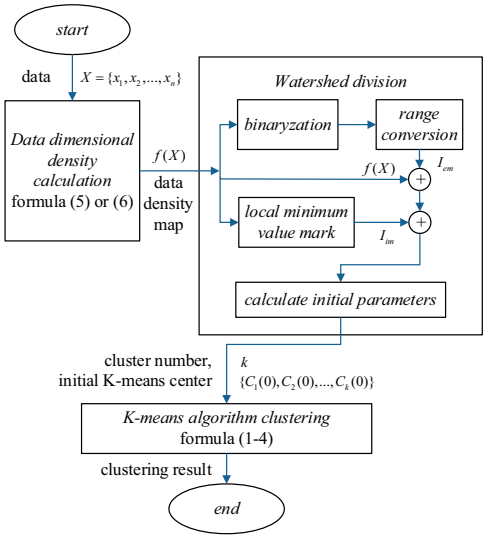


Fig. 2. Flow chart of the method

IV. CASE STUDY

The test data is shown in the following figure. This case is used for demonstration of the method operational process.

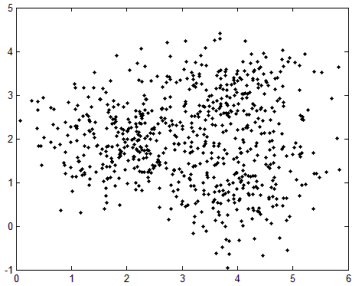


Fig. 3. Original data for clustering

This data set is produced by the two-dimensional normal distribution, whose probability density function shown as follow.

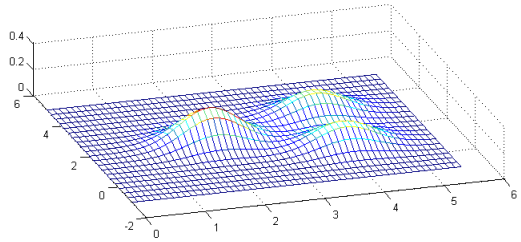


Fig. 4. Actual probability density distribution

According to (6), with parameter $\delta = 0.07$, the estimate density is shown in Figure 5.

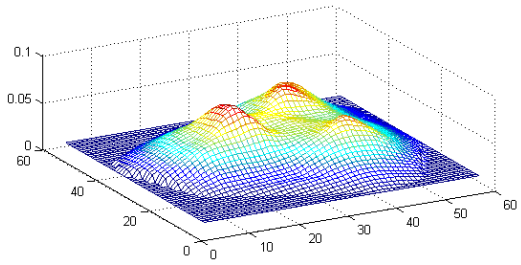


Fig. 5. Result of data relative dimensional density calculation

Put this figure upside down, and applied watershed method division. The result can be expressed as the following figure.

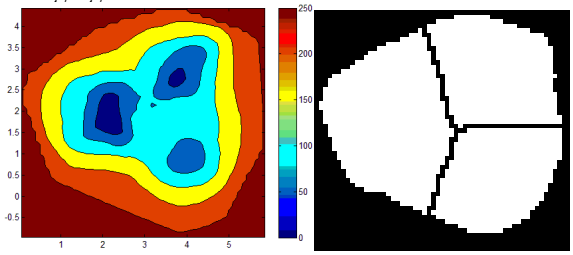


Fig. 6. Division after watershed method execution

The result shows that $k = 3$ is a more ideal cluster number. And through calculation, the initial K-means center is confirmed. The table below shows the relationship between estimated center, mean value of actual probability density distribution and clustering center after K-means algorithm.

TABLE I. RESULE COMPARATION

k=3	Class 1	Class 2	Class 3
estimated center	(1.768,1.721)	(4.169,2.783)	(4.225,0.551)
mean value of density distribution	(2.000,2.000)	(4.000,3.000)	(4.000,1.000)
center after K-means algorithm	(1.850,1.981)	(3.913,3.030)	(3.998,0.967)

The table explains that the estimated center is close to actual mean value of density distribution. And after K-means clustering, the result is much approximate to the real setting value.

The result of K-means clustering with parameters estimation is shown as following figure.

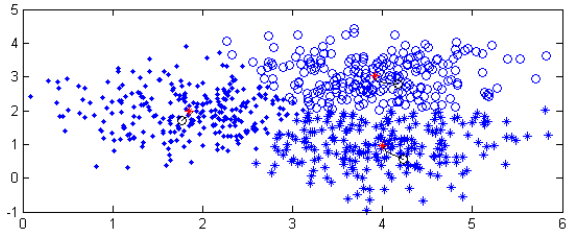
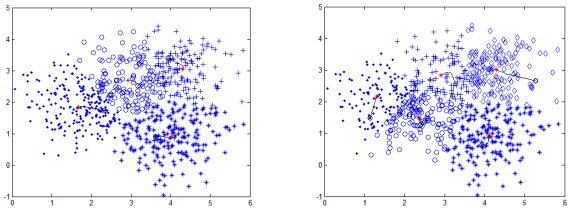


Fig. 7. Result of K-means clustering with parameters estimation

The red * in Fig. 7 represent the final clustering center, and black ○ stands for the initial estimated K-means center. Black lines between red * and black shows the variation of centers when K-means algorithm is iterating. From Table I and Fig.7, the result of such estimation is satisfactory.

Actually, without the initial parameter estimation, K-means algorithm can provided series of result according to different k given. The following figure showed two different situations without initial parameter estimation.



k=4 k=5

Fig. 8. Result with different k, but without parameters estimation

In order to study the influence of k value, 20 times of K-means clustering is done for different k. The result is shown in the following table.

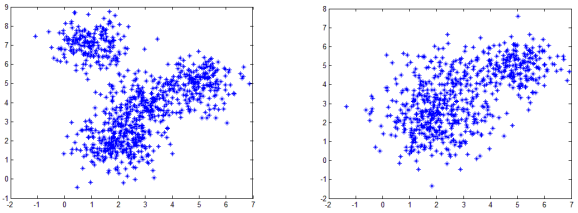
TABLE II. RESULE WITH DIFERERNT K VALUE

k	2	3	4	5	6
mean value of J[a]	1005.3	566.6	480.6	404.7	341.6
Variance of J	0.0	0.0	1009.3	269.4	689.6
average iteration times	11.3	9.1	16.1	17.6	19.6
k	7	8	9	10	
mean value of J	297.1	297.3	230.8	209.2	
Variance of J	681.1	147.7	832.8	342.7	
average iteration times	25.2	23.5	25.5	23.8	

[a]. J is calculated by Formula (2), and the mean value and variance refer to the 20 times test

It can be concluded that k=3 is suitable for the data set. and the average iteration time with parameter estimation is 8 times, which is also precede to the direct clustering method without the estimation.

Other data set would be used to test the estimation method in the following paragraph. These two data set shows the situation which data is too dense in distribution.



data set (a) data set (b)

Fig. 9. Original data for clustering

Applying the method discussed above, the results are shown as Fig.10, Fig.11 and Table III.

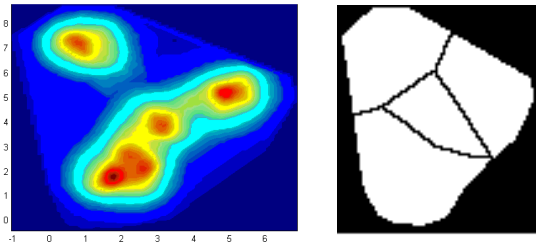


Fig. 10. Division after watershed method execution (for data set (a))

TABLE III. RESULE WITH DIFERERNT K VALUE

k=4	Class 1	Class 2	Class 3	Class 4
estimated center	(1.01,7.92)	(1.80,2.42)	(2.92,4.96)	(4.95,6.11)
mean value of density distribution	(1.00,7.00)	(2.00,2.00)	(3.00,4.00)	(5.00,5.00)
center after K-means algorithm	(1.0,7.06)	(1.99,1.86)	(2.86,3.97)	(5.01,4.99)

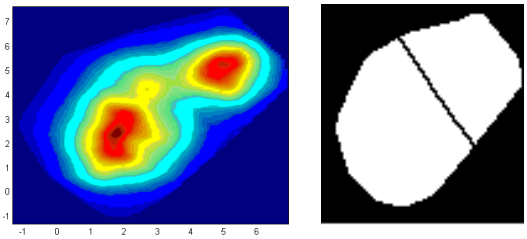


Fig. 11. Division after watershed method execution (for data set (b))

The method achieved good performance in data set (a), but in data set (b), the method divide the data into 2 clusters (the actually density distribution has 3 centers), which is on account of the data intensive.

V. CONCLUSION

This paper proposed an estimation of clusters number and initial centers of K-means clustering. The method is based on watershed method, which divides the data relative dimensional density distribution into multiple regions. Each regional center is selected as an initial K-means center, and the number of region is set as cluster number.

Case study shows the performance of such method is beneficial to the selection of K-means clustering initial parameter.

To operate K-means algorithm on large data set is less efficiency, but with initial parameter estimation, the algorithm enhanced both in veracity and running speed.

REFERENCES

[1] Li Jingjiao, Zhao Lihong, Wang Aixia. Pattern Classification. Beijing , China: Publishing House of Electronics Industry, 2010.

[2] Wei L, Zeng W, Wang H. “K-means clustering with manifold.” Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on, IEEE, 2010, pp:2095-2099.

[3] Kunhui Lin, Xiang Li, Zhongnan Zhang, Jiahong Chen. “A K-means Clustering with Optimized Initial Center Based on Hadoop Platform. ” The 9th International Conference on Computer Science & Education (ICCSE 2014) Aug 22-24, 2014. Vancouver, Canada, pp:263-266.

[4] Liu Guoli, Wang Tingting, YuLeimei, Li Yanping, Gao Jinqiao. “The Improved Research on K-Means Clustering Algorithm in Initial Values.” 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), Shenyang, China, Dec 20-22, 2013. pp:2124 – 2127.

[5] Shi Na, Liu Xumin, Guan Yong. “Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm.” Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, 2-4 April 2010. pp 63-67

[6] Zhang Z, Zhang Y. “Variable kernel density estimation based robust regression and its applications.” Neurocomputing, 2014, 134(9), pp:30–37.

[7] GE Shi-guo, WANG Mao-zhi, LIU Juan-juan, SU Jian-mei. “Improved Watershed Segmentation Algorithm Based on Marker. ” Journal of Sichuan University of Science & Engineering (Natural Science Edition). 2013,26(2), pp:57-61.

The Application of BP Neural Network Algorithm In Optical Fiber Fault Diagnosis

Shan Yan1
College of Science
North China University Of Science And Technology
*TangShan,China*e-mail: *64626027@qq.com*
Liu Yijuan2
JiTang College
North China University Of Science And Technology
TangShan,China
e-mail: 43129060 @qq.com

GUAN Fangjing3
School of Internet of Things Engineering
WuXi City College Of Vacational Technology
WUXI ,China
e-mail: guan_fj@163.com

Abstract –Adding the momentum is to build a suitable BP network model and the selected data are normalized to get the final sample data, and then analyze the data, by simulation experiments prove the feasibility that the BP neural network algorithm is applied in optical fiber fault diagnosis.

Keywords-BP Neural Network; Optical Fiber Fault Diagnosis

I. INTRODUCTION

With intelligent power grid put forward, electric power communication industry has an unprecedented period of development, optical fiber communication technology has become the first choice of modern power communication technology. There are many cable line faults, for example, the applications and the Long-time running of a large number of Electric power special optical fiber cables, aging and external damage that the cable is prone to. The faults are the key factors affecting the safe operation of electric power communication system. How to reduce or avoid optical fiber fault is the problem that all power communication workers has been committed to research.

Currently, BP (Back Propagation) neural network algorithm is one of the main research directions of data mining. Optical fiber fault diagnosis is mainly by obtaining a lot of light power data and using relevant research methods to reveal the relationship between Fiber optic cable running power alarm and the transmission situation, then provide the basis for making decision, in order to reduce loss caused by the fault. In this paper, BP neural network algorithm was improved by adding momentum; BP neural network technology is combined with power communication production, which is applied in optical fiber fault diagnosis. Available and valuable information hidden in the surface of the data is funded for managers to provide decision support, and it also is a new application of data mining.

II. THE IMPROVED BP NEURAL NETWORK ALGORITHM

BP algorithm is a feed-forward network algorithms role in promoting the use of the obvious, but the algorithm should be improved. Optimize the use of momentum here. Adding momentum is currently a more popular improved algorithm, the following is the formula of momentum weighting adjustment:

$$\omega(t+1) = -\eta \cdot \frac{\partial E}{\partial \omega} + \alpha \cdot V\omega(t) \qquad (1)$$

μ=0.9.

Improve the speed of network is directly determined by the momentum, this is because changing the value η. After such adjustment smoothed average direction toward the bottom of the variation. If the system enters the flat area, the error changes can be ignored, then $V\omega(t+1) \approx V\omega(t)$.

The average will change $V\omega = \frac{-\eta}{1-\alpha} \cdot \frac{\partial E}{\partial \omega}$。In the above formula $\frac{-\eta}{1-\alpha}$, the effectiveness of strong, can guarantee to get rid of velocity saturation region soon. Therefore, to accelerate the momentum in the learning process, but also to the convergence speed has been improved, while the realization of good results.

A. optimize the initial weights

Greatly affected the initial weights of the final result, even a direct result of the end result is good or bad, but in the process of correcting the weights, the selected algorithms greater impact on the convergence speed and precision. By the formula:

$$E^p = \sum_{k=1}^{q} (d_k^p - o_k^p)^2 = \sum_{k=1}^{q} (d_k^p - f(\sum_{j=1}^{m} \omega_{jk} y_j))^2 \qquad (2)$$

Recycling least squares method is determined in the hours:

$$\omega_{ij} = \frac{1}{\sum_{k=1}^{p} (\frac{x_{ij}}{y_{kj}})^2} \qquad (3)$$

The smaller the error, the weight it easier to meet the accuracy requirements, thus reducing the number of iterations.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.133

Meanwhile, in the process of initializing the weights, not just cycling debugging a set of data, but P sets of data, that exists between these data intrinsically linked, so the weight of the initialization process, it to all weighted, Set up the first floor of the right value is v_{ij} , at the same time satisfy the following formula:

$$\sum_{i=1}^{p} v_{ij} = 1, \forall j = 1,2,\cdots,n \quad (4)$$

As a result, the weight can successfully achieve uniformity of the sample.

B. Improved normalization method of training samples

Due to the comprehensive assessment of BP network input node physical quantity cannot do science comprehensive, because the difference between these quantities big, big gap on the numerical size, but there is no comparable place between training samples of each metric. Under normal circumstances, BP neural network model is based on the range [0, 1] S-type function as activation function. In this section, mainly quantified by calculation, the normalized initial data from one to [0.05,0.95], so that that the output can be retained for a long growth. The following equation (5) pre-processing of data, calculated using the actual sample, and all normalized to within [0.05, 0.95] range. Using the following formula processing normalization process:

$$\overline{X_i} = a + b \cdot \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

In the above formula: x_i represents the normalized before input and output; $\overline{X_i}$ represents normalized input and output; $x_{\min}$ 、 $x_{\max}$, Respectively, each set minimum and maximum factor variables.

C. determine the number of hidden layer

If you know the network level, so long as the increase to improve the accuracy of the hidden layer training it can be achieved. The number of the layers must be properly. According to equation (respectively nodes of the input layer and output layer, is constant between 1 and 10 to determine the number of hidden layer, choose the best result n, then you can achieve any of differentiable network function can be arbitrary precision infinitely close.

III. THE APPLICATION OF BP NEURAL NETWORK ALGORITHM IN OPTICAL FIBER FAULT DIAGNOSIS

Diagnosis fiber fault using BP neural network algorithm, first of all collecting optical power data, then followed by selection and cleaning, the data pretreatment, finally find out the useful information hidden.

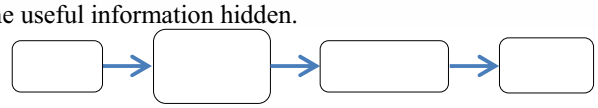


Fig.1 Data Mining Process of Optical Power

In this paper, based on the application of optical fiber monitoring system in Tangshan power supply company and the powerful function of Tangshan optical transmission network management, access to a large number of experimental data, applied in the improved BP neural network algorithm for simulation.

A. Data Collection

Data collection mainly is the optical power data collection, using SDH optical transmission network can be real-time query of the reception power value between two site emissions. Because of the continuous development of Tangshan power grid, four fiber loops select the only segment of cable which is not break or not T connection- from the city of Han to CheZhou Shan- as the data source of analysis and mining.

B. Data selection and cleaning

In the communication optical fiber cable, the numbers of the optical power change largely and are very unstable, which is influenced by many different environmental factors. Therefore, during the data collection, be sure to use the normal work of the equipment to collect light power data, otherwise some abnormal data because of large interference to the forecast ability of the network, the most serious result is to make the wrong decision by these data. For the study, the paper selected the optical power data of 6 days from the City of Han substation to the axle, and the data are used as the training sample data.

The optical power data gotten through communication transmission network have many properties, a series of information such as "Date/Time, OPR, opt, RS-BBE, RS-ES, RS-OFS, RS-SES" etc. However, not every information are useful for the data mining. Therefore, it is necessary to delete the data of useless information, to simplify the difficulty of data mining. The data after the preprocessing optical power information for data mining is shown in TableI.

TableI. Sheet Of Preprocessed Date

name	length	rem
OPR	4	0.1
OPT	3	0.1

C. Data Preprocessing

During the data processing, because the amplitude change data as output of BP neural network can cause great error, so the data should be standardized preprocessing, the distribution is in [0, 1].

S transfer function is mostly used in the process of normalization, the domain of function is [0, 1], and the function can be differential, characterized by showing a saturation nonlinearity. Based on above characteristics, the application of function in the nonlinear mapping ability of the network is very strong [2]. Its output curve is similar to the signal output of biological neurons, which are flat on the two sides, and the middle part is violent.

First, the input and output is normalized to [0, 1], then use the following formula to transformation.

$$\overline{X_i} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (6)$$

$\overline{X_i}$ is the data of input and output, $x_{\min}$ is the Minimum data transformation; x_{max} is Maximum data transformation..

Table Ⅱ The original data

OPT	OPR	OPT	OPR	OPT	OPR	OPT	OPR	OPT	OPR	OPT	OPR
-1.3	-3	-1.3	-3.6	-1.3	-7.1	-1.3	-3.1	-1.3	-3.5	-1.3	-3.6
-1.3	-3	-1.3	-3.4	-1.3	-11.	-1.3	-3.5	-1.3	-2.9	-1.3	-3.1
-1.3	-3.2	-1.3	-3.5	-1.3	-3.5	-1.3	-3.2	-1.3	-2.9	-1.3	-2.9
-1.3	-3.6	-1.3	-3.4	-1.3	-3.2	-1.3	-3.5	-1.3	-3.3	-1.3	-3.2
-1.3	-3.6	-1.3	-3.2	-1.3	-3	-1.3	-3.4	-1.3	-3	-1.3	-3.2
-1.3	-2.9	-1.3	-3	-1.3	-3.4	-1.3	-3.6	-1.3	-3.3	-1.3	-3.4
-1.3	-3.4	-1.3	-3.5	-1.3	-2.9	-1.3	-3.4	-1.3	-3.2	-1.3	-2.9
-1.3	-2.9	-1.3	-3.6	-1.3	-3.6	-1.3	-3.5	-1.3	-3	-1.3	-3
-1.3	-3.5	-1.3	-3.5	-1.3	-3.6	-1.3	-3.4	-1.3	-3.3	-1.3	-3.5
-1.3	-3	-1.3	-3.1	-1.3	-2.9	-1.3	-3.2	-1.3	-3.6	-1.3	-3.4
-1.3	-3.5	-1.3	-3	-1.3	-2.9	-1.3	-3	-1.3	-3.4	-1.3	-3.2
-1.3	-3.6	-1.3	-3.6	-1.3	-2.9	-1.3	-3.5	-1.3	-3	-1.3	-2.9
-1.3	-3.5	-1.3	-3.5	-1.3	-3.1	-1.3	-4	-1.3	-3.2	-1.3	-3.3
-1.3	-3.1	-1.3	-3.1	-1.3	-3.3	-1.3	-4.2	-1.3	-3.2	-1.3	-3
-1.3	-3	-1.3	-3	-1.3	-3	-1.3	-4.1	-1.3	-2.9	-1.3	-3
-1.3	-3.6	-1.3	-3.4	-1.3	-3.6	-1.3	-3.6	-1.3	-3.4	-1.3	-3.2

Through the calculation, can get two optical power attenuation values, and then normalized to get sample data.

TableⅢ The Sample Data

ATT	TRE	ATT	TRE	ATT	TRE	ATT	TRE	ATT	TRE	ATT	TRE
-1.7	0.9804	-2.3	0.9216	-5.8	0.5784	-1.8	0.9706	-2.2	0.9314	-2.3	0.9216
-1.7	0.9804	-2.1	0.9412	-9.8	0.1863	-2.2	0.9314	-1.6	0.9902	-1.8	0.9706
-1.9	0.9608	-2.2	0.9314	-2.2	0.9314	-1.9	0.9608	-1.6	0.9902	-1.6	0.9902
-2.3	0.9216	-2.1	0.9412	-1.9	0.9608	-2.2	0.9314	-2	0.951	-1.9	0.9608
-2.3	0.9216	-1.9	0.9608	-1.7	0.9804	-2.1	0.9412	-1.7	0.9804	-1.9	0.9608
-1.6	0.9902	-1.7	0.9804	-2.1	0.9412	-2.3	0.9216	-2	0.951	-2.1	0.9412
-2.1	0.9412	-2.2	0.9314	-1.6	0.9902	-2.1	0.9412	-1.9	0.9608	-1.6	0.9902
-1.6	0.9902	-2.3	0.9216	-2.3	0.9216	-2.2	0.9314	-1.7	0.9804	-1.7	0.9804
-2.2	0.9314	-2.2	0.9314	-2.3	0.9216	-2.1	0.9412	-2	0.951	-2.2	0.9314
-1.7	0.9804	-1.8	0.9706	-1.6	0.9902	-1.9	0.9608	-2.3	0.9216	-2.1	0.9412
-2.2	0.9314	-1.7	0.9804	-1.6	0.9902	-1.7	0.9804	-2.1	0.9412	-1.9	0.9608
-2.3	0.9216	-2.3	0.9216	-1.6	0.9902	-2.2	0.9314	-1.7	0.9804	-1.6	0.9902
-2.2	0.9314	-2.2	0.9314	-1.8	0.9706	-2.7	0.8824	-1.9	0.9608	-2	0.951
-1.8	0.9706	-1.8	0.9706	-2	0.951	-2.9	0.8627	-1.9	0.9608	-1.7	0.9804
-1.7	0.9804	-1.7	0.9804	-1.7	0.9804	-2.8	0.8725	-1.6	0.9902	-1.7	0.9804
-2.3	0.9216	-2.1	0.9412	-2.3	0.9216	-2.3	0.9216	-2.1	0.9412	-1.9	0.9608

ATT: Attenuation value, TRE: After treatment

D. BP Neural network training

The training of BP neural network is related to whether the network is successful, and the accuracy of BP network forecast is directly affected by the training. In the training results are not good, then we must improve and adjust the network structure, until meet the requirements of the test results, then you can use the trained network to analyze and forecast.

After the training of the neural network, need an important work-network test. During the process of test, use the data inconsistent with the training to test data [3-5]. The processes of BP network training and forecasting are different, they are two independent processes. After the success of the BP network training storage weights, these weights in the network forecasting will use.

E. Result analysis

When the optical transmission system is in application, during the transmission the attenuation value is generally less than 2.5dB, when the attenuation value is above 6dB, the system will make an alarm. If the value is between 2.5dB and 6dB for a long time, can determine to have some kind of defect or fault, at some time appear a rapid decay of the optical power, and ultimately lead to line fault.

As shown in Figure 2 for BP neural network prediction results, the solid line shows the optical power attenuation, the "·" is predictive value. There are many factors of influence for the prediction accuracy in the test, such as numbers processing after the decimal point of the data, and the accuracy of network model and training sample quantity, etc.

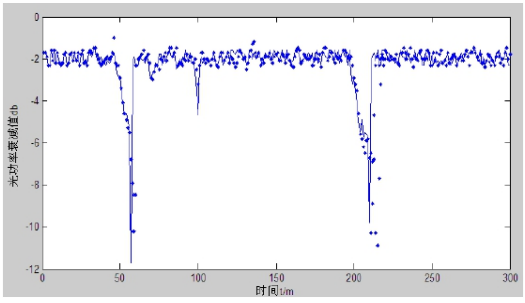


Fig.2 Simulation results

In Figure 2, you can intuitively see the situation is acceptable that a few of values deviates from the actual

attenuation value of the optical power, deviation of individual point minimally impact, and cannot make influence to the final results. So, the final results of the experiments and the result of the initial design t are consistent, that is, application of improved BP neural network algorithm in optical fiber fault diagnosis is feasible, the forecasting result is trusted and can provide valuable and available reference data for decision makers, and data mining is successful.

IV. CONCLUSION

In this paper, the BP algorithm is improved, which is applied to the mining and prediction of optical power data of electric power communication transmission network. In the actual network model, select the sample and construct training, and the process of its standardization uses BP. By constructing appropriate BP neural network and of the optical power value mining results are predicted and analyzed. It can not only help in the management and monitoring of cable, but also provide solid basis for the implementation of the decision-making.

[1] LAURA.J.NOVAK. Classification and prediction of stock price behavior[J]. Dissertation of Yale University, 2006(15):124-154. (references)

[2] MARTIN T. HAGAN, HOWARD B. DEMUTH, MARK BEALE. Neural Network Design[M]. Beijing: China Machine Press, 2002.

[3] He Xiaohui, Shi Rong. The method of Fault Diagnosis Based of BP Neural Network[J]. Computer and modernization, 2009(7):17-21.

[4] ZhangRui, Yang Xuanfang. Fault Diagnosis of Analog Circuits based on improved BP network[J]. Ordnance Automation, 2009, 28(9):71-73.

[5] Zhang Jun. Research on Visual Data Mining Technique[J]. Journal of Chongqing Technology and Business University(Science), 2013, 30(3):59-61.

.

2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Analysis Range of Coefficients in Learning Rate Methods of Convolution Neural Network

Jiang Zou, Qingbo Wu, Yusong Tan, Fuhui Wu, Wenzhu Wang
School of Computing, National University of Defense Technology
Changsha,China
Email:{zoujiang13@163.com} {qingbo.wu, yusong.tan, fuhui.wu, wenzhu.wang}@nudt.edu.cn

***Abstract*—Convolutional Neural Network (CNN) is a type of feed-forward artificial neural network, exploiting the unknown structure in input distribution to discover good representations with multiple layers of small neuron collections. CNN uses relatively little pre-processing compared to other classification algorithms, usually uses gradient decent to updates the parameters in the network. Since CNN was introduced in 1997s to deal with face recognition, it has made much achievement in many fields, and has been the state-of-the-art method in face recognition, speech recognition, etc. To get small error rate and a better speed of training the CNNs, a lot of learning rate methods are proposed. In this paper, we analyzed the range of the coefficients in these methods with a restriction of max convergence constant step learning rate. In our experiments, we find the max convergence learning rate by dichotomy with little computation cost, we also confirm the range of coefficients is useful. Moreover, we gives a comparison among these mothods on speed and error rate.**

***Keywords*-range of coefficients; max converge value; learning rate method; convolution neural network;**

I. Introduction

The Human Visual System (HVS) recognizes and localizes objects efficiently. Inspired by HVS, many Artificial Neural Networks are proposed to process the images, and CNN is one of them. CNN as the most promising architectures for images recognize by community consensus roughly mimics the nature of mammalian visual cortex. The hierarchical structure of CNN extracts localized features from input images, convolving images with filters. Following the filter responses are repeat sub-samplings and filters, then result in a deep feed-forward network architecture whose output feature vectors are classified. The capacity of CNN can be controlled by varying their depth and breadth. Compared to standard feed-forward neural networks with similar size layers, CNN has fewer connections and parameters due to the shared-weights, which means it is easier to train and takes fewer time.

CNN has made a great breakthrough in pattern classification, such as image recognition and speech recognition, after 2006. In the ILSVRC 2014, GoogleNet achieved a top-5 error rate of 6.67% in image classification. In view of the advantages of CNN and the high level of recognition rate, CNN is widely used. As the classified subjects and the datasets grow rapidly, the scale of CNN is getting bigger and the parameters of it is up to millions to billions. Despite the hardware factors, it is found that the computational speed is becoming a limiting factor for CNN.

In this paper, we are interested in the learning rate, which is the most important single parameter impacting the training speed and error rate. We list some on-hand learning rate methods and analysis the selection rules and the scope of the coefficients of these methods in theory by a limitation-the max convergence learning rate (η_m). Then we train a CNN with these methods to prove our conclusion. By dichotomy, we find the η_m in less than 600 iterations (which only need to do once for a dataset and a CNN), and then we get the value area of these coefficients in methods. Through our experiments, we find how we value the coefficients to get better performance on speed and lower error rate.

II. Convolutional Neural Netwrok

A. Structure of CNN

CNNs are hierarchical neural networks whose convolutional layers alternate with subsampling layers, reminiscent of simple and complex cells in the primary visual cortex [1][Wiesel and Hubel,1959]. The basic layers of CNN are image processing layer, convolutional layer, sub-pooling/Max-pooling layer, full-connected layer.

The image processing layer is an optional pre-processing layer, which adds information besides the raw input images, such as edges and gradients, with a predefined filters. Convolutional layer is used to get feature maps from the input with the convolution kernels, and it is parametrized by the size and the number of the maps, kernel sizes, skipping factors and the connection table. The parameters of each convolution kernel are trained by the backpropagation algorithm. The first convolution layers will obtain the lowest-level features, the more convolutional layers the network has , the higher-level features it will get.

Pooling layer is used to reduce variance, it computes the max or average value of a particular feature over a region of the image. Besides, the pooling layer also ensures that when there are small translations on the image features the same result will be obtained. The pooling layer in different architectural is different. In the CNN of [2][LeCun et al.1998] it is sub-sampling, but later it was found by [3][Scherer et al.,2010] that max-pooling can lead to faster convergence.

After the training in front layers, full-connected gets the high-level features as its input, then full-connected layer is used to classify these features.

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.134

B. Gradient Descent

In CNNs, we encounter the following optimization problem. For a given sequence of n training samples (x_1, y_1),..., (x_n, y_n), when $x_i \in R^d$,$y_i \in R$,CNN requires the solution of the following unconstrained optimization problem:

$$\text{Min } J(\omega) \text{ , } J(\omega) := \frac{1}{n}\sum_{i=1}^{n}\psi(\omega)$$

Where $\psi(\omega)$ is a loss function, for the loss function, we can use least squares regression, then

$$\psi(\omega) = (\omega^T x_i - y_i)^2$$

A standard update rule in a gradient descent can be describe as follows

$$\begin{aligned}\omega^{(t)} &= \omega^{(t-1)} - \eta_t \nabla J(\omega^{(t-1)}) \\ &= \omega^{(t-1)} - \frac{\eta_t}{n}\sum_{i=1}^{n}\nabla\psi(\omega^{(t-1)})\end{aligned}$$

Generally, we call η learning rate, and $\nabla J(\omega^{(t-1)})$ is the gradient of ω at iteration t . For simplicity, we use another equation to represent the update rule:

$$x_{t+1} = x_t + \nabla x_t \tag{1}$$

and

$$\nabla x_t = -\eta_t g_t \tag{2}$$

here, x_t represent the ω_t, g_t is the gradient of parameters.

C. learning rate methods

For all the hyper-parameters in the CNN, learning rate is the single most important one. Through different learning rate, the training converge or not, converge quickly or slowly, local optimum or overall optimum. So how to set learning rate is import to the training of CNN.

Constant step: The simplest way of scheduling η for training a CNN may be taking η as a constant. But not all the constant η can lead to converging. In our experiments, we find that if $\eta > \eta_m$ (η_m is the max converge value), the training will never converge. So the problem is how we get the critical value η_m and determine the η when $\eta < \eta_m$. Its hard and cost to find the critical value η_m by theoretical identification, and the critical value for one CNN or dataset may be too large or too small for another one. A common and useful method in practice is to pick different learning rate and test it in the training.

Reciprocal decrease: Another option is to decay the learning rate as the formula:

$$\eta(t) = b/(a+t) \tag{3}$$

In this formula a and b are constants, t is the iteration where parameters update. The value a, b decide the first value and the speed it decays. In this case, η begins with a relatively big value b/a and change rapidly to small values. It will results in slow convergence to bad solutions with a small b, and parameter blow-up for small t if b is too large (Darken and Moody, 1991, 1992, [4] [5])

STC: Search-then-converge(STC) is another better choice for the learning rate method (Darken and Moody, 1992, 1991, [4] [5]). In STC, η is chosen to be a fixed function of training iteration (n), and it was used in the CNN for face recognition in 1997 [6], in this paper, the function is as followed:

$$\eta(t) = \frac{\eta_0}{\frac{t}{N/2} + \frac{c_1}{max(1,(c_1 - \frac{max(0,c_1(t-c_2N))}{(1-c_2)N}))}} \tag{4}$$

Where η is initial learning rate, N is total training iterations, t is current training iteration, c_1 and c_2 are constant values.

Momentum method: Based on the constant step method, Momentum method was proposed in 1986. For the constant step method converge very slowly, the inclusion of a momentum term has been found to increase the rate of converge dramatically. With this method, the update rule is:

$$\nabla x_{t+1} = p\nabla x_t - \eta g_{t+1} \tag{5}$$

where p is the momentum parameter. That means, the modification of the weights at iteration $t+1$ depends on both the current gradient and the weights change of the previous iterations

ADAGRAD: A recent adaptive method called ADAGRAD [7] has shown remarkably good results on large scale learning tasks in distributed environment. The update rule for ADAGRAD is as follows:

$$\nabla x_t = -\frac{\eta}{\sqrt{\sum_{\tau=1}^{t} g_\tau^2}} g_t \tag{6}$$

Here η is a global learning rate shared by all dimensions, and the denominator computes all previous gradients on a per-dimension basis, So each dimension has its own dynamic rate. This method can be sensitive to initial conditions of the parameters and the corresponding gradients.

ADADELTA: Matthew presented a novel per-dimension learning rate method for gradient descent called ADADELTA in 2012. This method dynamically adapts over time using only first order information and has minimal computational overhead beyond vanilla stochastic gradient descent [8]. The update rule is:

$$E[g^2]_t = \rho E[g^2]_{t-1} + (1-\rho)g_t^2 \tag{7}$$

$$RMS[\nabla x]_{t-1} = \sqrt{E[\nabla x^2]_{t-1} + \epsilon} \tag{8}$$

$$RMS[g]_t = \sqrt{E[g^2]_t + \epsilon} \tag{9}$$

$$\nabla x_t = -\frac{RMS[\nabla x]_{t-1}}{RMS[g]_t} g_t \tag{10}$$

$$E[\nabla x^2]_t = \rho E[x^2]_{t-1} + (1-\rho)\nabla x_t^2 \tag{11}$$

$$\nabla x_{t+1} = x_t + \nabla x_t \tag{12}$$

This method implements previous squared gradients accumulation as an exponentially decaying average of the

squared gradient ($E[g^2]_t$)

III. ANALYSIS AND GUIDE FOR METHODS

When we get a CNN and train it with a learning rate method, what we want to know is how to select the coefficients of the method to ensure that the CNN training converge quickly and get a small error rate.

In this part, we analysis the range of coefficients of some on hands methods. Before we analysis the methods, we assume that these methods ensure the training is converged, so we assume that the loss function is decreasing substantially (although the actual training occasionally increase, it has little impact on the whole), then we get $|\nabla x_{t+1}| \leq |\nabla x_t|$ and $|g_{t+1}| \leq |g_t|$, that means $|\nabla x_t| \leq |\nabla x_1|$ and $|g_t| \leq |g_1|$.

Constant step: While we train the CNN with constant learning rate, it is simple but not efficient. First, we need to know the range of learning rate that make sure the training will converge; Second, we need to pick a constant which converge quickly and better to achieve a smaller error rate.

In this paper, we use dichotomy method as a quick search approach to find the max converge learning rate (η_m). During our search for η_m, we find that larger the η is (here $\eta_m \leq \eta$), more quickly the loss function value grows, the fewer iterations we need to judge whether the training is converging. After we test a η every time, we need to initial the parameters of the CNN to eliminate the impact of prior learning rate. Besides, we need to set a accuracy acc ($acc = |\eta_m - \eta|$) to stop the search to save the overhead.

Reciprocal decrease: For this method, learning rate begins with $\eta_1 = b/a$ and decays as t grows. To ensure the training will converge, we make sure: $b/a \leq \eta_m$. As update rule (3), b determines the rate of decline. With a larger b, it will decline more slowly, the training will converge more quickly. So when we chose the value of b, we can decide a with $a \geq b/\eta_m$

STC: From (4) we know STC method decay the η from η_0. compared to Reciprocal Decrease method, it decay more slowly. To ensure the training converge, we just need to satisfy $\eta_0 \leq \eta_m$.

Momentum method: Momentum method use the update rule(5).Here, we use $\eta_t = -\frac{\nabla x_t}{g_t}$ to represent the learning rate for per-dimension, in that way:

$$\begin{aligned}\eta_t &= -\frac{\nabla x_t}{g_t} \\ &= -\frac{\rho\nabla x_{t-1} - \eta g_{t+1}}{g_{t+1}} \\ &= \eta + \rho\frac{-\nabla x_{t-1}}{g_{t+1}}\end{aligned}$$

for $\nabla x_t \approx \nabla x_{t-1}$, we have

$$\eta_t \approx \eta + \rho\eta_t$$

to ensure the CNN converge, $\eta_t < \eta_m$,, then we have

$$\eta_t \approx \frac{\eta}{1-\rho} \leq \eta_m$$

at last, we get the coefficients range: $\rho \leq 1 - \frac{\eta}{\eta_m}$ and $\eta \leq (1+\rho)\eta_m$.

ADAGRAD: With a update rule (6), we have

$$-\frac{\nabla x_t}{g_t} = \frac{\eta}{\sqrt{\sum_{\tau=1}^{t} g_\tau^2}} \leq \frac{\eta}{|g_1|} \leq \eta_m$$

so the range of coefficient η is $\eta \leq \eta_m|g_1|$.

ADADELTA: The update rule of ADADELTA is (10), we use $\eta_t = -\frac{\nabla x_t}{g_t}$ to represent the learning rate for per-dimension. We assume that learning rate η_t is decreasing in the process of training (at least not increase), like that:

$$\begin{aligned}\eta_t &= \frac{RMS[\nabla x]_{t-1}}{RMS[g]_t} = \frac{\sqrt{E[\nabla x]_{t-1} + \epsilon}}{\sqrt{E[g^2]_t + \epsilon}} \\ &\leq \eta_1 = \frac{\sqrt{E[\nabla x]_0 + \epsilon}}{\sqrt{E[g^2]_1 + \epsilon}} = \frac{\sqrt{\epsilon}}{\sqrt{E[g^2]_1 + \epsilon}} = \frac{\sqrt{\epsilon}}{\sqrt{(1-\rho)g_1^2 + \epsilon}} \\ &\leq \eta_m\end{aligned}$$

then we have

$$\epsilon \leq \frac{\eta_m^2(1-\rho)g_1^2}{1-\eta_m^2}$$

usually, for $\eta_m^2 \ll 1$, we simplify the inequality as

$$\epsilon \leq \eta_m^2(1-\rho)g_1^2$$

at last, we get the range: $\epsilon \leq \eta_m^2(1-\rho)g_1^2$ and $\rho \leq 1 - \frac{\epsilon}{\eta_m^2 g_1^2}$.

IV. EXPERIMENT AND RESULTS

Data: We have used the CMU PIE face database, which consists of 3332 face images (68 individuals, 49 images for each). The images are grayscale with a resolution of 64x64 pixels. There are variations in facial expression (open/closed eyes, smiling/no smiling), facial details (glasses/no glasses), and different illumination intensity (weak illumination intensity to strong illumination intensity). For each individual, 35 images are chosen for training, and the rest 14 images are used for testing. Limited to the small dataset, after the training, we use all 3332 images as the test set and get the error rate.

CNN: We use a CNN model similar to the LeNet5, there are five layers: convolutional layer1, max-pooling layer1, convolutional layer2, max-pooling layer2, full connected layer. For the convolutional layer1, we get feature maps from the raw images through 5 convolution kernels with a size of 5x5, and we have 10 convolution kernels at the convolutional layer2. After the processing of convolution, we use max-pooling to get down-sampling of the feature maps pass from the front layers. For the two Pooling layers , the size of max-pooling is 2x2. For each such sub-region, output a maximum value. We set 2000 neurons at the full connected layer and 68 classifications.

MINI-batch: To employ vectorization and parallelism, CNN group training examples into mini-batches within SAG iterations in practice. In our experiment, we train 68 images for each iteration (35 iterations per epoch). So after one epoch, we have learnt all the images in the training image set.

A. Range of coefficients

In our experiment, we find the value of η_m by dichotomy (details can be seen in table 1). When we search η_m, larger the value is, smaller iterations the CNN oscillation, which means it is easier to judge whether the training converge. For example, it takes about 10 iterations to judge that it's not converging with $\eta = 0.5$, but it takes more than 200 iterations to judge training is not converging with $\frac{13}{256}$ ($\frac{13}{256} = 0.05078125$, where 0.05 ensure converge).

While we search η_m by dichotomy , we also recorded the g_t of each layers in CNN for several iterations. We find that the order of magnitude of g_t is different for different dimension in CNN. we estimates the order of magnitude of $|g_t|$ is $10^{-1} \sim 10^{-4}$. Because that larger the g_t is, the greater impact it has on the updating of parameters. To satisfy the convergence, we set the magnitude of g_t at 10^{-1}.

Table I
SEARCH η_m BY DICHOTOMY

η	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
converge	no	no	no	no	no	yes
oscillation	no	no	no	no	yes	yes
η	$\frac{3}{64}$	$\frac{7}{128}$	$\frac{13}{256}$	$\frac{25}{512}$	$\frac{51}{1024}$	$\frac{103}{2048}$
converge	yes	no	no	yes	yes	yes
oscillation	yes	yes	yes	yes	yes	yes

In order to facilitate the calculation, we set $\eta_m = 0.05$ to calculate the range of coefficients of methods.

Constant step: The range is: $\eta \leq 0.05$. we test different constant step on CNN (see table 2).

Reciprocal decrease: The range is $a \geq b\eta_m = \frac{b}{20}$. As a comparison, we also test whether training is converged when the coefficients exceed the range (see table 3).

STC: There are four coefficients in STC method: (η_0, c_1, c_2, N), but only η_0 decide whether the training is converging. The range of η_0 is $\eta_o \leq 0.05$.

Momentum method: There are two coefficients (ρ and η) in this method, as we calculated, their ranges are $\rho \leq 1 - 20\eta$ and $\eta \leq \frac{1+\rho}{20}$.

ADAGRAD: From the previous calculation we can know the range of coefficient is $\eta \leq \eta_m|g_1| = \frac{|g_1|}{20}$. In our experiment, the the scope of g_t we use to calculated coefficients range is [0.1,1], to calculate simple and easy to compare, we set $|g_t| = 0.2$, then $\eta \leq 0.01$

ADADELTA: The range of the coefficients of ADADELTA is $\epsilon \leq \frac{(1-\rho)g_1^2}{400}$ and $\rho \leq 1 - \frac{\epsilon}{400g_1^2}$. Because that the range of ϵ is positive correlated to g_1^2, and g_t has a wide range of values, it's difficult to obtain precise ranges of ϵ, For simplify, we use estimated value. For example, while $\rho = 0.99$, we set $g_1 = 0.2$, then $\epsilon \leq 10^{-6}$; while $\rho = 0.90$, we set $g_1 = 0.2$, then $\epsilon \leq 10^{-5}$.

B. Discussion

Constant step: As shown in table 2, when train with η_m, it wins one of the fastest speeds. Overall, the error rate doesn't have a certain link with the value of η.

Table II
CONSTANT STEP

η	converge	iterations	error rate
0.01	yes	2205	3.39%
0.02	yes	2695	1.68%
0.03	yes	1634	2.10%
0.04	yes	1584	1.50%
0.05	yes	939	2.20%
0.06	no	–	–

Table III
RECIPROCAL DECREASE

b/(a+t)	in scope	convg	iters	error rate
1/(10+t)	no	yes	> 5000	–
1/(20+t)	yes	yes	> 5000	–
10/(10+t)	no	no	–	–
10/(100+t)	no	no	–	–
10/(200+t)	yes	yes	1429	3.69%
100/(1000+t)	no	no	–	–
100/(2000+t)	yes	yes	1120	3.96%

Table IV
SEARCH-THEN-CONVERGE

(η_0, c_1, c_2, N)	in scope	convg	iters	error rate
(0.1,50,0.65,1000)	no	no	–	–
(0.07,50,0.65,1000)	no	yes	683	2.67%
(0.05,50,0.65,1000)	yes	yes	1062	3.54%
(0.04,50,0.65,1000)	yes	yes	1215	3.96%

Table V
MOMENTUM METHOD

(ρ, η)	in scope	convg	iters	error rate
(0.90, 0.01)	no	no	–	–
(0.80, 0.01)	yes	no	–	–
(0.73, 0.01)	yes	yes	1715	1.35%
(0.70, 0.01)	yes	yes	1593	1.56%
(0.80, 0.02)	no	no	–	–
(0.70, 0.02)	no	no	–	–
(0.60, 0.02)	yes	yes	980	7.59%
(0.50, 0.02)	yes	yes	1266	1.6%
(0.60, 0.03)	no	no	–	–
(0.50, 0.03)	no	no	–	–
(0.40, 0.03)	yes	no	–	–
(0.35, 0.03)	yes	yes	713	4.14%
(0.30, 0.03)	yes	yes	980	2.31%

Table VI
ADAGRAD

η	in scope	convg	iters	error rate
0.10	no	no	–	–
0.05	no	no	–	–
0.01	yes	no	–	–
0.005	yes	yes	1348	4.02%
0.001	yes	no	–	–

Table VII
ADADELTA

(ρ, ϵ)	in scope	convg	iters	error rate
$(0.99, 10^{-6})$	yes	no	–	–
$(0.99, 10^{-7})$	yes	yes	858	1.44%
$(0.99, 10^{-8})$	yes	yes	1184	1.44%
$(0.99, 10^{-9})$	yes	yes	2409	1.11%
$(0.95, 10^{-5})$	yes	no	–	–
$(0.95, 10^{-6})$	yes	yes	980	0.90%
$(0.95, 10^{-7})$	yes	yes	1266	1.50%
$(0.95, 10^{-8})$	yes	yes	2246	2.16%
$(0.95, 10^{-9})$	yes	yes	3594	2.19%
$(0.90, 10^{-5})$	yes	no	–	–
$(0.90, 10^{-6})$	yes	yes	1062	1.38%
$(0.90, 10^{-7})$	yes	yes	1674	0.99%
$(0.90, 10^{-8})$	yes	yes	2164	2.70%
$(0.90, 10^{-9})$	yes	yes	3839	2.73%
$(0.80, 10^{-4})$	no	no	–	–
$(0.80, 10^{-5})$	yes	yes	1552	0.87%
$(0.80, 10^{-6})$	yes	yes	1266	1.14%
$(0.80, 10^{-7})$	yes	yes	2205	1.23%
$(0.80, 10^{-8})$	yes	yes	2369	2.76%
$(0.80, 10^{-9})$	yes	yes	4206	2.85%
$(0.70, 10^{-4})$	no	no	–	–
$(0.70, 10^{-5})$	yes	yes	858	0.84%
$(0.70, 10^{-6})$	yes	yes	1144	0.81%
$(0.70, 10^{-7})$	yes	yes	2205	0.63%
$(0.70, 10^{-8})$	yes	yes	3349	1.50%
$(0.70, 10^{-9})$	yes	yes	5309	2.31%
$(0.60, 10^{-4})$	no	no	–	–
$(0.60, 10^{-5})$	yes	yes	1225	0.81%
$(0.60, 10^{-6})$	yes	yes	1225	1.08%
$(0.60, 10^{-7})$	yes	yes	2491	0.78%
$(0.60, 10^{-8})$	yes	yes	4574	0.87%
$(0.60, 10^{-9})$	yes	yes	7000	1.65%

Reciprocal decrease: In order to get a better speed, except for η_m, we also need to know the excepted training iterations. Due to the shared learning rate, it doesn't achieve a good performance on error rate.

STC: The table 4 shows an exceptional case: when begin to decay with 0.07, it is not only converged, but also achieve a good training speed. That means a slightly larger value will accelerate the training, but it will still converge if we decay the learning rate appropriately.

Momentum method: Our experimental results and theoretical calculations have some errors, some value in scop result in non-converge, but these errors are within the normal range. From the table 5 will find that with a larger ρ and smaller η, it will takes more iterations, but it will get smaller error rate.

ADAGRAD: When we training with ADAGRAD we find that it's not a easy job to choose a suitable η. Due to denominator of the learning rate is sum of g_t^2, learning rate will drop rapidly when the train starts (decay to 10^{-4} or smaller in dozens of iterations), at the same time, great oscillation will occurs, that means it may be never converged.

ADADELTA: Because of the good performance on error rate, we made a detailed test on this method. From table 7 we can see that this method perfectly match our theoretical value range. When we value the coefficients with critical value (for example, $(0.95, 10^{-6})$), it ends training in 980 iterations and achieve an error rate of 0.9% which never appears in other methods in our experiments. What's more, ADADELTA has a wide range of value for coefficients ensuring convergence, which brought us great convenience to pick a pair of (ρ, ϵ).

V. Conclusion

In this paper, we analysis the selection rules and the range of coefficients of different learning rate methods which used in CNN training. We test with different value of coefficients and the results prove that the range of coefficients we calculated works well in the training. The experiments also give a clear comparison among the different methods. There is no doubt that the ADADELTA method wins the best comprehensive performance on speed, error rate, stability.

References

[1] Hubel D H, Wiesel T N. Receptive fields of single neurones in the cat's striate cortex[J]. The Journal of physiology, 1959, 148(3): 574-591.

[2] Lecun Y, Bottou L. Gradient-Based Learning Applied to Document Recognition, in PROC,1998.

[3] Scherer D, Mller A, Behnke S. Evaluation of pooling operations in convolutional architectures for object recognition[M]//Artificial Neural NetworksCICANN 2010. Springer Berlin Heidelberg, 2010: 92-101.

[4] Darken C, Moody J. Towards faster stochastic gradient search[C]//NIPs. 1991: 1009-1016.

[5] Darken C, Moody J. Note on learning rate schedules for stochastic optimization[R]. YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE, 1992.

[6] Lawrence S, Giles C L, Tsoi A C, et al. Face recognition: A convolutional neural-network approach[J]. Neural Networks, IEEE Transactions on, 1997, 8(1): 98-113.

[7] J. Duchi, E. Hazan, and Y. Singer, Adaptive subgradient methods for online leaning and stochastic optimization, in COLT, 2010.

[8] Zeiler M D. ADADELTA: An adaptive learning rate method[J]. arXiv preprint arXiv:1212.5701, 2012.

Improved Feature Selection Based On Normalized Mutual Information

LI Yin[1] , Ma Xingfei[1], Yang Mengxi[1]
1: Education Information Research Center
WuXi Vocation Institute of Commerce
Wuxi, China
e-mail:liyin@wxic.edu.cn

Zhao Wei[1], Gu Wenqiang[2]
2: Department I, Software Division, Shanghai Huaqin
Telecom Co.Ltd,
Shanghai, China
E-mail:gu.wenqiang@163.com

Abstract—For the question (NMIFS) algorithm has the disadvantages of redundancy. This paper introduces a new feature selection method by enhanced NMIFS algorithm. A new quality estimation function is introduced in the new feature selection algorithm to overcome the shortcomings of the classic NMIFS, and the experiment shows on that normalized mutual information feature selection The experiment shows that the INMIFS can generate impressive results in accuracy and redundancy.

Keywords-Mutual Information (MI), feature selection algorithms, Normalized Mutual Information Feature Selection (NMIFS)

I. INTRODUCTION

A feature selection algorithm can be used to classify the feature subsets which are identified and removed as much of the irrelevant and redundant information as possible, along with an evaluation measure. The best subset contains the least number of dimensions that most contributed to accuracy. The feature selection is important to speed up training and to improve generalization performance[1].

In this active field of research, numerous classic feature selection algorithms have been widely-used, such as wrappers, filters and embedded methods[2].

Filter methods use a measure to capture the usefulness of the feature subsets from the high-dimension data sets, for example, using the common measures which based on the mutual information, it can allow the feature selection algorithms to operate faster and more effectively.

The traditional feature selection algorithms use Shannon's mutual information (MI) as a measure of relevance among features. But the MI method has the disadvantages of redundancy. In 1994, Battiti [3] proposed mutual information feature selection (MIFS) which selected the feature that maximizes the information about the class, corrected by subtracting a quantity proportional to the average MI with the previously selected features. Kwak and Choi [4] analyzed the limitations of MIFS and proposed a greedy selection method called MIFS-U, which in general, makes a better estimation of the MI between input attributes and output classes than MIFS. The average normalized mutual information feature selection (NMIFS) was introduced by Estevez [5]in 2009, it propose an initialization procedure and a mutation operator based on NMIFS to speed up the convergence of the genetic algorithm.

II. MUTUAL INFORMATION

Mutual Information [6] is used to quantitatively analyze the mutual dependence between any two features or between a feature and a class variable.

The MI of two continuous random variables X and Y can be defined as:

$$I(X;Y) = \iint_{XY} P(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy \qquad (1)$$

Where $P(x), P(y), P(x,y)$ are the marginal probability distribution functions of X,Y, and (X,Y). Considering to two discrete random variables X and Y, the MI can be defined as:

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.135

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2(\frac{p(x,y)}{p(x)p(y)}) \quad (2)$$

Additionally, the mutual information can also be represented by the entropy. It can be equivalently expressed as

$$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X) \quad (3)$$

Where H(X) and H(Y) are the marginal entropy, The joint entropy of X and Y is related to the conditional entropy.

$$H(X|Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2 p(x|y) \quad (5)$$

$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y) \quad (6)$$

Thus, $$I(X;Y)=H(X)+H(Y)-H(X,Y) \quad (7)$$

Where two discrete random variables X and Y, the MI of X and Y can be defined as

$$I(X,Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2(\frac{p(x,y)}{p(x)p(y)}) \quad (8)$$

Note that the MI of two random variables X and Y is 0, it illustrates X and Y are generally independent of each other, and there are no correlation. However, the higher the dependency of two variables, the greater the value of mutual information (MI).

III. IMPROVED NORMALIZED MUTUAL INFORMATION FEATURE SELECTION ALGORITHM （INMIFS）

A. Normalized Mutual Information Feature Selection Algorithm（NMIFS）

In 2009, Estevez proposed the normalized mutual information feature selection. In NMIFS algorithm, for the initial feature sets fi and fs, the MI definition can be rewritten in terms of entropy and conditional entropy as follows:

$$I(f_i;f_s)=H(f_i)-H(f_i|f_s)=H(f_s)-H(f_s|f_i) \quad (9)$$

From Eq. (9), the MI can take values in the following interval:

$$0\le I(f_i;f_s)\le \min\{H(f_i),H(f_s)\} \quad (10)$$

$H(x|y),H(y|x)$ are the conditional entropy.

$$H(X)=-\sum_{x\in X}p(x)\log_2 p(x) \quad (4)$$

Then, we can define the normalized MI between *fi* and *fs* with the minimum entropy of both features.

$$NI(f_i;f_s)=\frac{I(f_i;f_s)}{\min\{H(f_i),H(f_s)\}} \quad (11)$$

From eq.(11), we can find $NI(f_i,f_s)\in[0,1]$. A value 0 indicates that feature and the subset of selected features are independent. A value 1 indicates that feature is highly correlated with all features in the subset. In order to reduce the limitation of redundancy between feature subsets. We defined the evaluation function J(f) as follows,

$$J(f)=I(C;f)-\frac{1}{|S|}\sum_{s\in S}NI(f_i;f_s) \quad (12)$$

Where C is the Class variable, S is the subset of selected features, |S| is the cardinality of the set S. In process of NMIFS, we choose the feature f that has the maximizes measure J(f) as the next feature.

NMIFS algorithm has more effectively, and accuracy on feature selection can be improved. However, the evaluation function still has the disadvantages of redundancy between feature subsets in some cases. For example, assume that the feature f just has an obvious relation with one feature value of the set S at redundancy, and has smaller relation with other feature values of the set S. Then it makes the value of $\frac{1}{|S|}\sum_{s\in S}\frac{I(f;s)}{\min\{H(f),H(s)\}}$ smaller, and still chooses this feature f as the selected features. In actual, there are many redundancies in the subset selected features S, which has bad effect in the dependability of the data.

B. Improved Normalized Mutual Information Feature Selection Algorithm（INMIFS）

Based on relevance and redundancy[7], we should consider fully that the relation between the candidate feature f and each feature of the set S. In this paper, we propose a novel feature selection by using the following alternative feature quality estimation function.

$$J(f) = I(C;f) - \max\left\{\frac{I(f;s)}{\min\{H(f),H(s)\}}\right\} \quad (13)$$

This selection criterion introduces the maximum value between the candidate feature and the set of selected features instead of the average normalized MI[8] as above. The work-flow of INMIFS algorithm is illustrated as follows,

(1) Initialize：Set F as the initial set of N features, and set S as an empty set.

$F = \{f_1, f_2, \cdots f_i, \cdots f_N\}$, $1 \le i \le N$, and $S = \{\Phi\}$

(2) Compute the MI: According to the Eq.(8),for each feature f_i ($f_i \in F$), calculate the $I(C;f_i)$.

(3) Select the first feature: Find the feature f_i that maximizes $I(C;f_i)$, except f_i from F, and then set $S = \{f_i\}$.

(4) Greedy selected: repeat until $|S| = k$:

(4.1) Compute the MI between variables: Calculate $I(f_i, f_s)$ for all pairs (f_i, f_s), with $f_i \in F$ and $f_s \in S$, if it is not already available.

(4.2) Select the next feature: Choose the feature f_i that maximizes *J(f)* via Eq.(13) as the selected features. Except f_i from F, and then set $S = S \bigcup \{f_i\}$.

(5) Output the set S containing the selected feature.

In INMIFS algorithm, we calculate the MI between the feature and the class variable ($I(C;f_i)$), the computational cost is $O(N \log N)$. Meanwhile, in step (4) process, the complexity is $O(KN \log N)$, when computing the selected features.The total of complexity is the same as NIMIFS ($O(N \log N)$)[9].

IV. EXPERIMENTAL RESULTS

A. Dataset

We use MATLAB 7.9 as a development platform and complete experiment in the Windows XP operating system.The three real datasets[11] (segmentation, krvs, and vehicle) are downloaded from the UCI machine learning repository. Table 1 shows summary of the datasets.

TABLE I. SUMMARY OF THE DATASETS

Dataset	No. of sets	Size	No. of classes
Segmentation	19	2310	7
Krvs	36	3196	2
Vehicle	18	946	4

Note that the following feature sets (region-pixel-count, short-line-density-5 and short-line-density-2) have the same values in Segmentation dataset. They have no effect on the results of cluster. We only choose the other sets for effective feature selection.

B. Results

In this paper, the No. of selected features, the classification accuracy, and the training performance with Bayesian and Decision Tree are used to evaluate the classifying results and to compare the quality of the feature selections algorithms.

(1) The No. of selected features

TABLE II. THE NO. OF SELECTED FEATURES FOR INMIFS AND NMIFS

Dataset	No. of sets	NMIFS	INMIFS
Segmentation	19	11	7
Krvs	36	21	12
Vehicle	18	14	12

From the above table, it can remove more irrelevant

features from the initial set via INMIFS algorithm. In special, it is a better choice to use INMIFS in high-dimensional vectors, such as Krvs dataset.

(2) The classification accuracy

TABLE III. THE CLASSIFICATION ACCURACY IN DECISION TREE CLASSIFIER（%）

Dataset	Full Set	NMIFS	INMIFS
Segmentation	95.9294+0.0104	96.8831+0.0104	96.9697+0.0101
Krvs	91.3261+0.0105	95.025+0.0835	95.9324+ 0.0653
Vehicle	72.4586+0.1415	73.1678+0.1441	74.5863+0.1366

TABLE IV. THE CLASSIFICATION ACCURACY IN BAYESIAN CLASSIFIER（%）

Data	Full Set	NMIFS	INMIFS
Segmentation	80.2165+0.0575	83.7662+0.0534	83.7662+0.0534
Krvs	87.8911+0.0397	90.5194+0.1567	91.612+ 0.245
Vehicle	44.7991+0.2826	45.6265+0.2912	44.6898+0.296

In the above tables, compared with no feature selection algorithm, the accuracy of NIMFS and INMIFS in Classifier are higher. The INMIFS is a promising feature selection method in terms of highest accuracy.

(3) The training performance

We utilize Mat-lab Arsenal toolbox with WEKA [12] integrated to implement our training experiments. Therefore, we provide a more general training view of the Decision Tree (DT), Bayesian (BS) classification methods.

From the following figures, compared with NMIFS algorithm, the INMIFS algorithm has significant improvement in classification performance. In the beginning, the accuracy is higher with the increasing number features, and then it would stabilize when reaching the certain number features. We obtain the best selected features number. Thus, the INMIFS algorithm ensures the lower redundancy.

ACKNOWLEDGMENT

This paper proposes a feature selection method by enhanced NMIFS algorithm. A new quality estimation function is introduced in the INMIFS algorithm to overcome the shortcomings of the classic NMIFS, and it generates impressive results in accuracy and redundancy. This algorithm has been adopted and applied successfully by an data analysis based on campus card system in Wuxi vocation institute of commerce.

REFERENCES

[1] John G H, Kohavi R, P flegerK. Irrelevant Features and the Subset Selection Problem. Proc of the 11th International Conference on Machine Learning. New Brunswick, USA, 1994: 121 – 129

[2] Huang J, Cai Y, Xu X. A hybrid genetic algorithm for feature selection wrapper based on mutual information [J]. Pattern Recognition Letters, 2007, 28: 1825-1844.

[3] Battiti R. Using mutual information for selecting features in Supervised neural net learning [J]. IEEE Transactions on Neural Networks. 1994. 5 : 537-550.

[4] Kwak N, Choi C.H. Input feature selection by mutual information based on Parzen window[J].IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002, 24(12):1667-1671 .

[5] Estevez P A, Tesmer M, Perez C A, et al. Normalized Mutual Information Feature Selection [J].IEEE Transactions on Neural Networks, 2009, 20(2): 189-201.

[6] Fatemeh Amiri, Mohammad Rezaei. Mutual information-based feature selection for intrusion detection systems [J].Journal of Network and Computer Applications,34(2011)1184-1199.

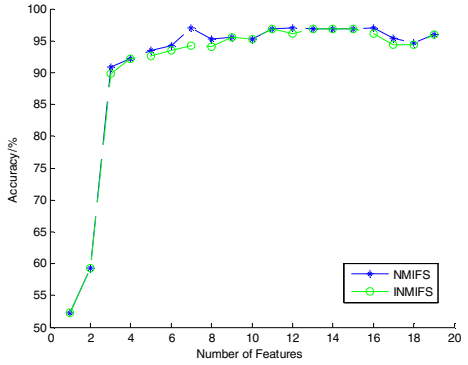
[7] Petar Ristoski, Heiko Paulheim, Feature Selection in Hierarchical Feature Spaces, Discovery Science Lecture Notes in Computer Science Volume 8777, 2014, pp 288-300

[8] N. Hoquea, D.K. Bhattacharyyaa, J.K. Kalitab, A mutual information-based feature selection method, Expert Systems with Applications Volume 41, 2014, Pages 6371 – 6385

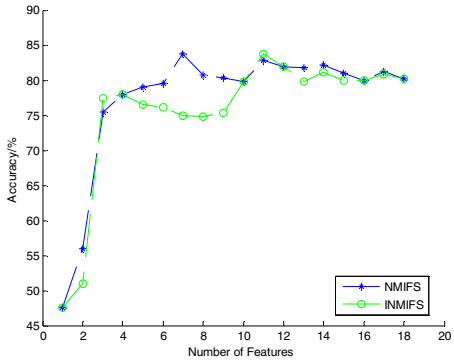
[9] Peng H, Long F, Ding C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[10] Blake C, Merz C. UCI repository of machine learning database[EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.

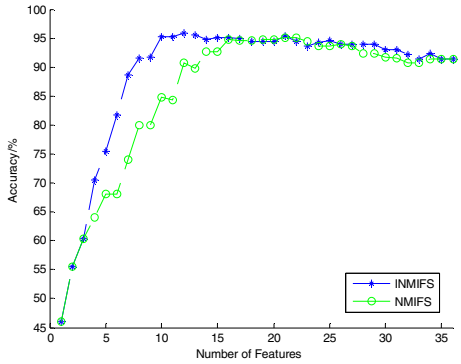
[11] Witten I H, Frank E. Data Mining Practical Machine Learning Tools and Techniques with JAVA Implementations[M].2nd ed.CA: Morgan Kaufmann,2005.



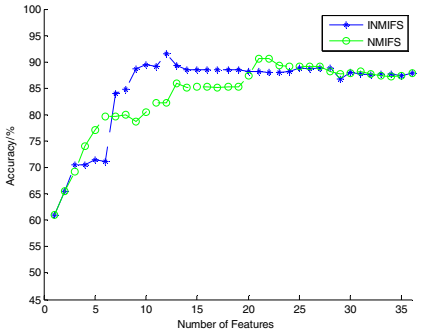
（a）Segmentation+ Decision Tree



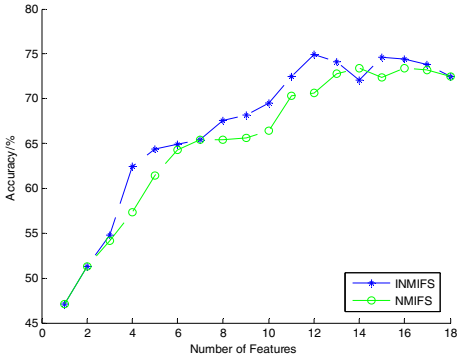
（b）Segmentation+ Bayesian



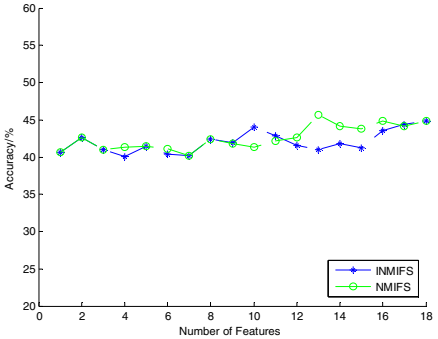
（c）Krvs + Decision Tree



（d）Krvs + Bayesian



（e）Vehicle+ Decision Tree



（f）Vehicle+ Bayesian

Fig1. The training views of the DT, BS classification with INMIFS and NMIFS

A modified Dynamic Window Approach to Obstacle Avoidance combined with Fuzzy Logic

Zhang Hong[1, 2*], Sun Chun-long[2], Zheng Zi-jun[2], An Wei[1, 2], Zhou De-qiang[1, 2], Wu Jing-jing[1,2]

[1]Jiangsu Provincial Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi, 214122, China

[2]College of Mechanical Engineering, Jiangnan University, Wuxi, 214122, China

E-mail: jndxzh@qq.com

* Corresponding author

***Abstract*—As a classic reactive method for obstacles avoidance, the dynamic window approach (DWA) uses an objective function to choose the optimal velocity commands. The objective function consists of three weighted terms, respectively urging the robot to face to the target point, avoid crash and move fast. The environment where robot works is usually complex and changeable, and different situations need different weights. However, the original DWA and its extensions use constant weights, and there is no existing algorithm for changing them. In this paper, we propose a modified DWA combined with fuzzy logic. This can be done by analyzing the information about goal and obstacles, and then using fuzzy logic to decide suitable weights in real time. In contrast with the original DWA, the proposed method makes robot move more safely and smoothly.**

***Keywords*—*Local Path Planning*, *Dynamic Window Approach*, *Weights*, *Fuzzy Logic*.**

I. INTRODUCTION

Local path planning is an important part of autonomous motion of mobile robot, which asks the robot to move toward the preferred location without any crash in a completely or partially unknown environment. In recent years, lots of methods for obstacle avoidance have been proposed, and among them the dynamic window approach (DWA) [1] is famous because it takes the dynamic properties of the robot into account.

DWA is first proposed by D. Fox in 1997, and it performs well on the robot named RHINO. Compared with other obstacle avoidance methods, DWA concerns the constraints of the robot such as maximum velocity, acceleration, etc. Hence, the control command gained from this method is seemed more in line with the actual situation. Some extensions of this approach have also been published in the past years. O. Brock applied DWA in global obstacle avoidance problem in [2], and C. Schoröter combined particle filter with DWA to optimize objective function in [3]. P. Saranrittichai took the obstacles which are near the trajectory but not on it into account and modified the objective function to make robot move more safely [4].

Concerning than the environment around robot is usually complex and changeable, M. Seder took FD* search algorithm into DWA to do obstacle avoidance in dynamic environment [5]. B. Choi used fuzzy logic to change the speed of robot according to the obstacle density [6]. These extended methods modified the objective function consisting of weighted terms more or less, but never thought to change the weights. D. Kiss even proposed a global dynamic window-based navigation method without weighted objective function [7].

Generally speaking, different situations need different weights. In this work, we proposed a modified DWA combined with fuzzy logic. The modified method analyzes the information about goal and obstacles, and then uses fuzzy logic to decide suitable weights in real time.

This paper is organized as follows. In Section II, we give a brief introduction of the original DWA. The detail of the modified DWA is presented in Section III. Simulation and experimental results are shown in Section IV. Finally, we draw a conclusion to the proposed method in Section V.

II. DYNAMIC WINDOW APPROACH

A. Introduction Of Dynamic Window Approach

As is mentioned above, Dynamic Window approach is a classic method for obstacle avoidance which is first proposed by D. Fox in [1]. The concept of DWA is to get an optimal velocity command in the velocity space directly by maximizing the objective function. To achieve that, there are two steps in DWA. Details of each step are as follows:

The first step is to limit the search space. The velocity command is gained from the velocity space of the robot. There are three kinds of velocity space defined in DWA and each velocity in the velocity space is a tuple (v, ω), in which v is the translational velocity and ω is the rotational velocity. The first velocity space is made up of all the possible velocities the robot can achieve. Each velocity in the second velocity space ensures that the robot can stop before crashing any obstacle if we choose this velocity. The third velocity space contains the velocities can be reached in a clock cycle, and it is also called the Dynamic Window. If these three velocity spaces can be expressed as V_s, V_a and V_d in order, then the resultant velocity search space is given by:

$$V_r = V_s \cap V_a \cap V_d \tag{1}$$

The second step is to pick the most appropriate velocity command. Firstly, each velocity in V_r will be sampled to calculate its appropriateness with objective function. In original DWA, objective function is defined as follows:

$$G(v,\omega) = a \cdot heading(v,\omega) + b \cdot dist(v,\omega) + c \cdot vel(v,\omega) \tag{2}$$

978-1-4673-6593-2/15 $31.00 © 2015 IEEE
DOI 10.1109/DCABES.2015.136

, where a, b and c are constant, and usually $a + b + c = 1$.

In this function, $heading(v, \omega)$ maintains the course of the robot towards its target point; $dist(v, \omega)$ measures the closeness of the robot with the closest obstacle for any sampled velocity; $vel(v, \omega)$ is to force the robot to move fast.

Finally, the velocity command for the next time cycle is the velocity with the highest value of the objective function. That is to say

$$V_{cmd} = ArgMax\{G(v,\omega)\} \tag{3}$$

B. Disadvantage Of Dynamic Window Approach

Using original DWA and its extensions, the robot can perform obstacle avoidance toward target point successfully and safely in many simple cases. However, these methods always use a constant set of weights, while the environment around robot is usually complex and changeable. As a result, the success rate decreases as the obstacle density increases and the surrounding becomes crowded.

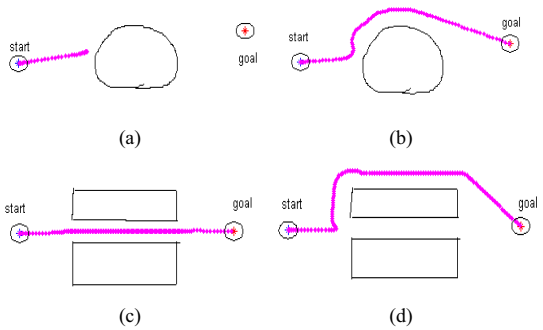


Fig. 1 Example of local path planning when robot tries to move round an obstacle or pass through the passageway using DWA with constant weights: in (a) and (c) $a = 0.8$, $b = 0.1$, $c = 0.1$; in (b) and (d) $a = 0.1$, $b = 0.8$, $c = 0.1$.

Different values of a, b and c result in different paths when robot moves. The set with low value of a and high value of b gives the robot much freedom in moving round obstacles, as shown in Fig. 1(b). At the same time it may stop the robot from passing through passageway between two obstacles (Fig. 1(d)). On the other extreme, the set with high value of a and low value of b leads robot to go through the narrow passageway (Fig. 1(c)). If so, it may force the robot to approach objects very closely before turning away. In Fig. 1(a), the robot even stops before obstacle because DWA chooses (0, 0) to be the optimal velocity.

There is no set of weights is universal, and the simplest solution is to change the values of the weights in real time, making the objective function more adaptive to current environment. In the next section, we use fuzzy logic to change the set of weights in real time according to the environment around the robot.

III. The Proposed Method

The objective function of DWA contains three parts: target heading, obstacle avoidance and fast moving. That is to say, changing the values of the weights means changing the priorities of these three parts. Fuzzy logic is often used in similar problem. Yang Jing-dong also divides the local path plan into three kinds of behavior, and use fuzzy logic to decide the weight of each kind, and finally fuse them [8]. In this paper, we use fuzzy logic to change the set of weights in real time according to the distribution of obstacles and the position of target.

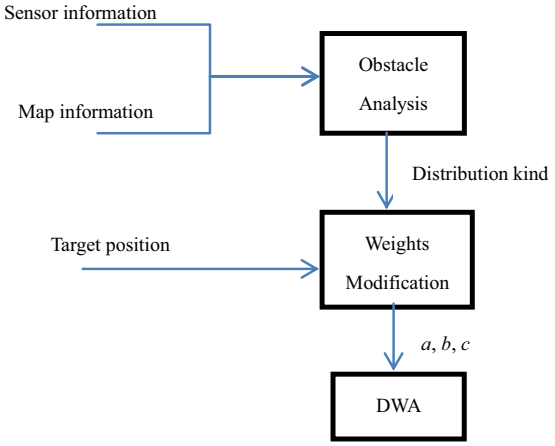


Fig. 2 Procedure of proposed method

The procedure of proposed method is shown in Fig. 2. Before robot chooses a trajectory to avoid the obstacles around, information about the obstacles is gained from sensors and known map. According to the information, the distribution of obstacles is recognized by the first fuzzy controller and then sent to the second fuzzy controller together with information about the target point. Finally, the output of the fuzzy controller is the velocity command that we need. The details of the distributions of obstacles and fuzzy rules are defined as follows:

A. The distributions of obstacles

Usually there are eight kinds of distribution of obstacles, as shown in Fig. 3. In order to recognize current kind of distribution, we set the range of distance obstacle to be [0, 2m], and the fuzzy language is described as {near, far}. The fuzzy inputs are obstacle distances of three directions of robot, and the fuzzy output is the kind of distribution, fuzzy language is discretely described as {a ~ h}. For example, if the left distance is near and the left distance is near and the left distance is near then the kind of distribution would be h.

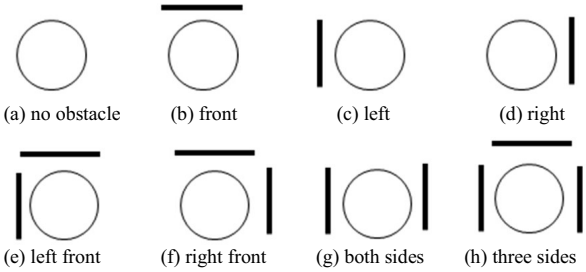


Fig. 3 Kinds of distribution of obstacles around robot

B. *Modifying weights in real time*

To modify the weights in real time, we design a fuzzy controller with three inputs and three outputs. Three inputs are the kind of distribution of obstacles, the distance from robot to target point and the angle of the target point relative to the robot's heading direction, expressed as K_o, d_t and θ_t in order. The distance from robot to target point ranges in [0, 10m], described as {N, M, F}; the angle of the target point relative to the robot's heading direction ranges in [0, 180°], described as {Small, Mid, Large}. The three outputs are the weights of the objective function in DWA, whose range is [0, 1], described as {VS, S, M, H, VH}.

When obstacles are in the front of robot, DWA should be given a higher obstacle avoiding weight b; when target point is near and the angle is large, DWA should be given a higher target heading weight a; when the environment around robot is empty, DWA can be given a higher fast moving weight c. Knowing that, the fuzzy rule can be set as follows:

If K_o is a and d_t is F and θ_t is Small, then a is VS and b is VS and c is VH;

If K_o is b and d_t is F and θ_t is Small, then a is VS and b is H and c is S;

If K_o is c and d_t is F and θ_t is Small, then a is VS and b is S and c is H;

…

If K_o is h and d_t is N and θ_t is Large, then a is VH and b is VH and c is VS;

Fig. 4 Fuzzy rules of proposed method

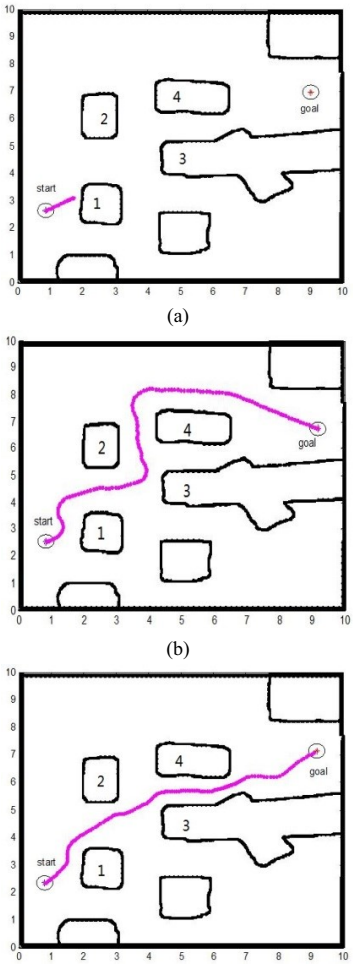
We use Mamdani Inference to do fuzzy reasoning. After getting the exact output values of weights, we should normalize them to meet $a + b + c = 1$.

IV. Simulation and Experimental Results

A. *Simulation results*

We simulate our method in Matlab, and the map size is 10m×10m. The max translational velocity is 1m/s and translational acceleration is 1m/s^2. The max rotational velocity is 90°/s and the rotational acceleration is 90°/s^2. The robot shape is assumed to be circular shape and the radius of the robot is 0.25m. The clock cycle is set 0.2s, meaning that DWA will find a control command every 0.2 seconds. However, the interval to modify weights is set 1s, because it costs a relatively long time to do fuzzy reasoning.

In Fig. 5(a), we set the weights to be (0.6 0.3 0.1), but the robot still stop before the first obstacle as similar as that shown Fig. 1(a). In Fig. 5(b), we set the weights to be (0.3 0.6 0.1), and the robot success to move round the first obstacle and pass through the wider passageway between obstacle 1 and 2. However, when robot tries to move round obstacle 3, it is unable to enter the passageway between obstacle 3 and 4 but choose a longer way. Finally in Fig. 5(c), we apply the proposed method to modify the weights, the path of robot is much better than the first two, showing an advantage over the original DWA.



(a)

(b)

(c)

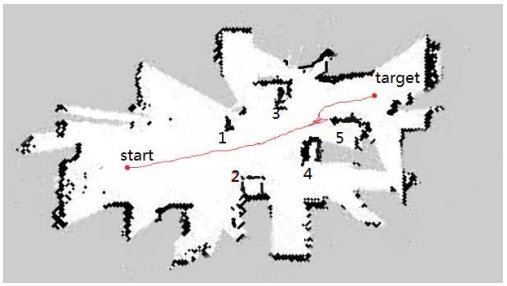
Fig. 5 Simulation result. The original DWA with different sets of weights are used in (a) and (b); the proposed method.is used in (c).

B. *Experimental results*

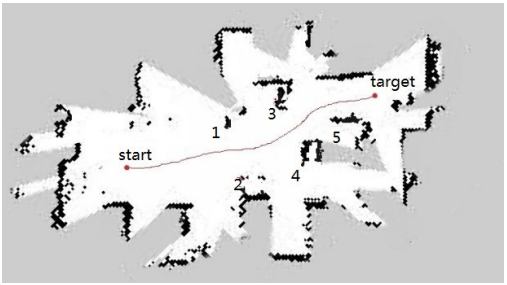
The experiment is performed on the Turtlebot robot (Fig. 6(a)), which is equipped with a Kinect sensor to measure the obstacle data from the environment. The program is written in C++ on Robot Operation System, running on Laptop CPU Core i5-2430 2.4 GHz. Fig. 6(b) shows the experimental environment, and the robot need to move from the start point to the target point.

(a) TurtleBot (b) Experimental environment

Fig. 6 Experimental setting



(a) The original DWA



(b) The proposed method

Fig. 7 Experimental results

We perform the experiment with the original DWA and the proposed method, respectively. We choose (0.6 0.3 0.1) to be the weights in the first experiment, and the result is shown in Fig 7(a). The robot passed through obstacle 1, 2, 3 and 4 successfully and smoothly. When the robot tries to move round the obstacle 5, it crashes. This is because that the target heading weight *a* is relatively high, which results in clearance between obstacle 5 and robot. After three attempts of 'forward - backward' moving, the robot succeeds to reach the target point.

Then we perform the second experiment with the proposed method. As shown in Fig 7(b), the robot can modify the weighs of DWA in real time. Therefore, when the robot moves to the target position, it has a relatively safe distance to the obstacles as well as a smooth and continuous moving path.

V. Conclusions

In this work, we propose a modified Dynamic Window Approach combined with fuzzy logic to do local path planning. In the proposed method, the weights of objective function are modified in real time by using fuzzy logic so as to make the DWA more adaptive to current environment around robot. Simulation and experimental results shows its advantage over the original DWA.

References

[1] D. Fox, W. Burgard and S. Thrun, "The Dynamic Window Approach to Collision Avoidance," IEEE Robotics & Automation Magazine, vol. 4, no. 1, pp. 23-33, 1997.

[2] O. Brock and O. Khatib, "High-speed Navigation Using the Global Dynamic Window Approach," IEEE International Conference on Robotics & Automation, May, 1999, pp. 341–346.

[3] C. Schröter, M. Höchemer and H. M. Gross, "A Particle Filter for the Dynamic Window Approach to Mobile Robot Control," Proceeding of 52nd International Scientific Colloquium, 2007, vol. 1, pp. 425-430.

[4] P. Saranrittichai and N. Niparnan, "Robust Local Obstacle Avoidance for Mobile Robot based on Dynamic Window Approach," Proceeding of 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunication and Information Technology (ECTI-CON), 2013, pp. 1-4.

[5] M. Seder and I. Petrović, "Dynamic window based approach to mobile robot motion control in the presence of moving obstacles," IEEE International Conference on Robotics and Automation, April, 2007, pp. 10-14.

[6] B. Choi, B. Kim, E. Kim. "A modified Dynamic Window Approach in crowded indoor environment for intelligent transport robot," Proceeding of 12th International Conference on Control, Automation and Systems, 2012, pp. 1007-1009.

[7] D. Kiss, G. Tevesz, "Advanced dynamic window based navigation approach using model predictive control," Proceeding of 17th International Conference on Methods and Models in Automation and Robotics (MMAR), 2012, pp. 149-153.

[8] Yang Jing-dong, Yang Jing-hui, "Obstacle Avoidance Based on Multiple Objective Optimization for Mobile Robots," Journal of Shang Hai Jiao Tong University, vol. 46, no. 2, pp. 213-216, 2012.

Author Index

Author Index

Author Index

Author Index

IEEE Computer Society Technical & Conference Activities Board

IEEE Computer Society Publications

The world-renowned IEEE Computer Society publishes, promotes, and distributes a wide variety of authoritative computer science and engineering texts. These books are available from most retail outlets. Visit the CS Store at *http://www.computer.org/portal/site/store/index.jsp* for a list of products.

IEEE Computer Society *Conference Publishing Services* (CPS)

The IEEE Computer Society produces conference publications for more than 300 acclaimed international conferences each year in a variety of formats, including books, CD-ROMs, USB Drives, and on-line publications. For information about the IEEE Computer Society's *Conference Publishing Services* (CPS), please e-mail: cps@computer.org or telephone +1-714-821-8380. Fax +1-714-761-1784. Additional information about *Conference Publishing Services* (CPS) can be accessed from our web site at: *http://www.computer.org/cps*

Revised: 18 January 2012

CPS Conference Publishing Services — CPS Online™ IEEE Online Collaborative Publishing Environment

CPS Online is our innovative online collaborative conference publishing system designed to speed the delivery of price quotations and provide conferences with real-time access to all of a project's publication materials during production, including the final papers. The ***CPS Online*** workspace gives a conference the opportunity to upload files through any Web browser, check status and scheduling on their project, make changes to the Table of Contents and Front Matter, approve editorial changes and proofs, and communicate with their CPS editor through discussion forums, chat tools, commenting tools and e-mail.

The following is the URL link to the ***CPS Online*** Publishing Inquiry Form:
http://www.computer.org/portal/web/cscps/quote