Lecture Notes in Electrical Engineering 274

James J. (Jong Hyuk) Park
Hojjat Adeli
Namje Park
Isaac Woungang *Editors*

Mobile, Ubiquitous, and Intelligent Computing

MUSIC 2013

Springer

Lecture Notes in Electrical Engineering

Volume 274

For further volumes:
http://www.springer.com/series/7818

James J. (Jong Hyuk) Park · Hojjat Adeli
Namje Park · Isaac Woungang
Editors

Mobile, Ubiquitous, and Intelligent Computing

MUSIC 2013

Springer

Editors
James J. (Jong Hyuk) Park
Department of Computer Science and Engineering
Seoul University of Science and Technology
Seoul
Republic of Korea

Hojjat Adeli
Departments of Biomedical Informatics, Civil and Environmental Engineering and Geodetic Science, and Neuroscience, and Centers of Biomedical Engineering and Cognitive Science
Ohio State University
Columbus
Ohio
USA

Namje Park
Department of Computer Education
Teachers College
Jeju National University
Korea

Isaac Woungang
Department of Computer Science
Ryerson University
Toronto, Ontario
Canada

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-3-642-40674-4 ISBN 978-3-642-40675-1 (eBook)
DOI 10.1007/978-3-642-40675-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013947701

Message from the MUSIC 2013 General Chairs

MUSIC 2013 is the FTRA 4th International Conference on Mobile, Ubiquitous, and Intelligent Computing (MUSIC 2013). This conference takes place September 4–6, 2013, in Gwangju, Korea. The aim of the MUSIC 2013 was to provide an international forum for scientific research in the technologies and application of Mobile, Ubiquitous and Intelligent computing. The MUSIC-13 is the next edition of the 3rd International Conference on Mobile, Ubiquitous, and Intelligent Computing (MUSIC-12, Vancouver, Canada, 2012) which was the next event in a series of highly successful International Workshop on Multimedia, Communication and Convergence technologies MCC-11 (Crete, Greece, June 2011), MCC-10 (Cebu, Philippines, August 2010).

The papers included in the proceedings cover the following topics: Mobile Computing, Ubiquitous Computing, Intelligent Computing, Intelligent and Mobile Services. Accepted and presented papers highlight new trends and challenges of Information Technology Convergence and its Services. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research. We would like to express our sincere thanks to Steering Chairs: James J. (Jong Hyuk) Park (SeoulTech, Korea). Our special thanks go to the Program Chairs: Namje Park (Jeju National University, Korea), Jinjun Chen (Swinburne University of Technology, Australia), Isaac Woungang (Ryerson University, Canada), Song Guo (University of Aizu, Japan), Soon Ae Chun (City University of New York, USA), all Program Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

September 2013

Sang Soo Yeo
Hojjat Adeli
Qun Jin
Fionn Murtagh

MUSIC 2013 General Chairs

Message from the MUSIC 2013 Program Chairs

Welcome to the 4th International Conference on Mobile, Ubiquitous, and Intelligent Computing (MUSIC 2013) will be held in Gwangju, Korea, September, 4–6, 2013. The MUSIC-13 will be the most comprehensive conference focused on the various aspects of Mobile, Ubiquitous and Intelligent computing. The MUSIC 2013 will provide an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of intelligent technologies in mobile and ubiquitous computing environment. In addition, the conference will publish high quality papers which are closely related to the various theories, modeling, and practical applications in MUSIC. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in this important subject.

For MUSIC 2013, we received many paper submissions, after a rigorous peer review process, we accepted 36 articles with high quality for the MUSIC 2013 proceedings, published by the Springer. All submitted papers have undergone blind reviews by at least two reviewers from the technical program committee, which consists of leading researchers around the globe. Without their hard work, achieving such a high-quality proceeding would not have been possible. We take this opportunity to thank them for their great support and cooperation. We would like to thank all authors for their work and presentation, all members of the program committee and reviewers for their cooperation and time spent in the reviewing process. Particularly, we thank the founding steering chair of MUSIC 2013, James J. (Jong Hyuk) Park. Finally, Special thanks are extended to the staffs from MUSIC 2013, who contributed so much to the success of the conference.

Thank you and enjoy the conference!

Namje Park, Jeju National University, Korea
Jinjun Chen, Swinburne University of Technology, Australia
Isaac Woungang, Ryerson University, Canada
Song Guo, University of Aizu, Japan
Soon Ae Chun, City University of New York, USA

MUSIC 2013 Program Chairs

Organization

Steering Chairs

James J. Park	SeoulTech, Korea

General Chairs

Sang Soo Yeo	Mokwon University, Korea
Hojjat Adeli	The Ohio State University, USA
Qun Jin	Waseda University, Japan
Fionn Murtagh	Royal Holloway, University of London, UK

Program Chairs

Namje Park	Jeju National University, Korea
Jinjun Chen	Swinburne University of Technology, Australia
Isaac Woungang	Ryerson University, Canada
Song Guo	University of Aizu, Japan
Soon Ae Chun	City University of New York, USA

Advisory Committee

Seok Cheon Park	Gachon University, Korea
Makoto Takizawa	Seikei University, Japan
William I. Grosky	University of Michigan-Dearborn, USA
Young-Sik Jeong	Dongguk University, Korea
Mohammed Atiquzzaman	University of Oklahoma, USA
Leonard Barolli	Fukuoka Institute of Technology (FIT), Japan
Timothy K. Shih	National Central University, Taiwan
V. Clincy	Kennesaw State University, USA
Wenny Rahayu	La Trobe University, Australia

Sethuraman Panchanathan	Arizona State University, USA
Dimitrios G. Stratogiannis	National Technical University of Athens (NTUA), Greece

Workshop Chairs

Deok Gyu Lee	ETRI, Korea
Chin-Feng Lee	Chaoyang University of Technology, Taiwan
Dakai Zhu	University of Texas at San Antonio, USA

Publicity Chairs

Jen-Wei Hsieh	National Taiwan Univ. of Sci. and Tech., Taiwan
Eric Renault	Institut Telecom - Telecom SudParis, France
Wanqing Tu	Nottingham Trent University, UK

Publication chair

Hwa-Young Jeong	Kyung Hee University, Korea

Program Committee

Ahmad Khasawneh	Hashemite University, Jordan
Alagan Anpalagan	Ryerson University, Canada
Alejandro Linares Barranco	Universidad de Sevilla
Alvaro Marco	Universidad Zaragoza, Spain
Basit Shafiq	Lahore University of Management Sciences, Pakistan
Bratislav Milic	Humboldt-Universitat zu Berlin, Germany
Changjing Shang	Aberystwyth University, UK
Chee-Peng LIM	Deakin University, Australia
Cherie Ding	Ryerson University, Canada
Chih-Heng Ke	National Kinmen Institute of Technology, Taiwan
Chin-Feng Lai	National Ilan University, Taiwan
Dooho Choi	Electronics and Telecommunications Research Institute, Korea
Heechang Shin	Iona College, USA
Hongxin Hu	Delaware State University, USA
Jaime Lloret Mauri	Universidad Politecnica de Valencia, Spain
Jaya Parvathy	SSN College of Engineering, Chennai, India
Jose Luis Sevillano Ramos	Universidad de Sevilla, Spain
Jun Xiao	East China Normal University, China
Junichi Suzuki	University of Massachusetts, USA

Kami Makki	Lamar University, TX, USA
Kassem Saleh	Kuwait University, Kuwait
Kok-Seng Wong	Soongsil University, Korea
Kyoil Cheong	Electronics and Telecommunications Research Institute, Korea
Liang Zhou	Nanjing University of Posts and Telecommunications, China
Ling-Jyh Chen	Academia Sinica, Taiwan
M. Elena Renda	IIT-CNR, Italy
Marco Vieira	University of Coimbra, Portugal
Mehul Bhatt	University of Bremen, Germany
Michel Barbeau	Carleton University, Canada
Muhammad Younas	Oxford Brookes University, UK
Narayanan Kulathuramaiyer	Universiti Malaysia Sarawak, Malaysia
Natalia Stakhanova	University of South Alabama, USA
Nik Bessis	University of Derby, UK
Oliver Amft	Eindhoven University of Technology, Potentiaal
Pascal Lorenz	University of Haute Alsace, France
Peter Mueller	IBM Zurich Research Laboratory, Switzerland
Petros Nicopolitidis	Aristotle University, Greece
Qiang Zhu	The University of Michigan, USA
Qinghe Du	Xi'an Jiaotong University, China
Rachid Anane	Coventry University, UK
Raj Jain	Washington University in St. Louis, USA
Rami Yared	Japan Advanced Institute of Science and Technology, Japan
Sanjay Kumar Dhurandher	University of Delhi, India
Shao-Shin Hung	WuFeng University, Taiwan
Shi Zhefu	University of Missouri, Kansas, USA
Soohyun Oh	HOSEO UNIVERSITY, Korea
Sung-Hyuk Cha	Pace University, USA
Susan Imberman	CUNY College of Staten Island, USA
Taesung Kim	KISA, Korea
Tat-Chee Wan	Universiti Sains Malaysia, Malaysia
Tian Tian	Manhattan University, USA
Wanqing Tu	Glyndwr University, UK
Wendy Hui Wang	Stevens Institute of Technology, USA
Wenhua Wang	Marin Software, USA
Wenjia Li	Georgia Southern University, USA
Xavier Fernando	Ryerson University, Canada
Yoo Jung An	Essex County College, USA
Young soo Kim	Electronics and Telecommunications Research Institute, Korea
Youngsook Lee	Howon University, Korea
Zhou Su	Waseda Univ., Japan

Contents

Active Media Technologies (AMT)

A Novel Ranking Technique Based on Page Queries 1
Gwangbum Pyun, Unil Yun

Ranking Book Reviews Based on User Discussion 7
Heungmo Ryang, Unil Yun

The Blog Ranking Algorithm Using Analysis of Both Blog Influence and Characteristics of Blog Posts 13
Jiwon Kim, Unil Yun

Frequent Graph Mining Based on Multiple Minimum Support Constraints 19
Gangin Lee, Unil Yun

Design of Automatic Paper Identification System with QR Code for Digital Forensics 25
Ha-Kyung Jennifer Lee, Young-Mi Yun, Kee-Hyung Yoon, Dong-Sub Cho

Computational Awareness for Telecommunication /Energy-Efficient Systems

Processing Continuous Range Queries with Non-spatial Selections 31
HaRim Jung, Seongkyu Kim, Joon-Min Gil, Ung-Mo Kim

DSPI: An Efficient Index for Processing Range Queries on Wireless Broadcast Stream 39
Kwanho In, Seongkyu Kim, Ung-Mo Kim

End-to-End High Speed Forward Error Correction Using Graphics Processing Units 47
Md Shohidul Islam, Jong-Myon Kim

Human Centric Computing and HCI

DirectSpace: A Collaborative Framework for Supporting Group Workspaces over Wi-Fi Direct 55
Jong-Eun Park, Jongmoon Park, Myung-Joon Lee

Internet of Things (IoT)

Specification of Communication Based Train Control System Using AADL 63
Lichen Zhang, Bingqing Xu

An Agent Modeling for Overcoming the Heterogeneity in the IoT with Design Patterns 69
Euihyun Jung, Ilkwon Cho, Sun Moo Kang

BK-means Algorithm with Minimal Performance Degradation Caused by Improper Initial Centroid 75
Hoon Jo, Soon-cheol Park

Portable and Smart Devices

Detect Spatial and Temporal Gait Parameters by Dual Accelerometers 81
Wann-Yun Shieh, An-Peng Liu, Tyng-Tyng Guu

Implementation of Load Management Application System in Energy Management Service 87
Taekyeong Kang, Hyungkyu Lee, Dong-Hwan Park, Hyo-Chan Bang, Namje Park

Toward a Mobile Application for Social Sharing Context 93
Meng-Yen Hsieh, Ching-Hung Yeh, Yin-Te Tsai, Kuan-Ching Li

A Research Based on the Effect of Smart Phone Use on Consumption Life of Teenagers in a Smart Era 99
Jeonghan Son, Keon Uk Kim, Yeon-gyeong Seo, Wonyeong Oh, Seowon Choi, Ana Kang

Security and Monitoring of WSN

Quality-Workload Tradeoff in Pig Activity Monitoring Application 105
Haelyeon Kim, Yeonwoo Chung, Sungju Lee, Yongwha Chung, Daihee Park

Security, Privacy, Authentication, Trust for IoT

Applying Different Cryptographic Algorithms for Mobile Cloud Computing 111
Sung-Min Jung, Nam-Uk Kim, Seung-Hyun Lee, Dong-Young Lee, Tai-Myoung Chung

Intrusion-Tolerant Jini Service Architecture for Ensuring Survivability of U-Services Based on WSN 117
Sung-Ki Kim, Jae-Yeong Choi, Byung-Gyu Kim, Byoung-Joon Min

Creation Mechanism for Access Group Based on User Privacy Policy-Based Protection 125
Taekyeong Kang, Hyungkyu Lee, Dong-Hwan Park, Hyo-Chan Bang, Namje Park

Ubiquitous Context Awareness

Specification of Train Control Systems Using Formal Methods 131
Bingqing Xu, Lichen Zhang

Formal Descriptions of Cyber Physical Systems Using Clock Theory 137
Bingqing Xu, Lichen Zhang

An Intelligent Dynamic Context-Aware System Using Fuzzy Semantic Language 143
Daehyun Kang, Jongsoo Sohn, Kyunglag Kwon, Bok-Gyu Joo, In-Jeong Chung

WSN Applications and Technologies

Efficient Data Monitoring in Sensor Networks Using Spatial Correlation 151
Jun-Ki Min

Power-Time Tradeoff of Parallel Execution on Multi-core Platforms 157
Sungju Lee, Heegon Kim, Yongwha Chung

Effective Object Identification through RFID Reader Power Control 165
Shung Han Cho, Sangjin Hong, Nammee Moon

Market-Based Resource Allocation for Energy-Efficient Execution of Multiple Concurrent Applications in Wireless Sensor Networks 173
Mo Haghighi

Clustering Objects in Heterogeneous Information Network Using Fuzzy C-Mean 179
Muhammad Shoaib, Wang-Cheol Song

Data-Intensive Intelligence and Knowledge

Better Induction Models for Classification of Forest Cover 185
Hyontai Sug

Application for Temporal Analysis of Scientific Technology Information 191
Myunggwon Hwang, Do-Heon Jeong, Jinhyung Kim, Jangwon Gim, Sa-kwang Song, Sajjad Mazhar, Hanmin Jung, Shuo Xu, Lijun Zhu

ROI Extraction in Dermatosis Images Using a Method of Chan-Vese Segmentation Based on Saliency Detection 197
Zehan Wang, Lijun Zhu, Jiandong Qi

The Study on Semantic Self-sufficiency in Factual Knowledge Extraction 205
Yunliang Zhang

XML-Based Document Retrieval in Chinese Diseases Question Answering System 211
Haodong Zhang, Lijun Zhu, Shuo Xu, Weifeng Li

Mathematical Document Retrieval Model Using Structural Information of Equations in Pseudo-documents 219
Yeongkil Song, Junsoo Shin, Harksoo Kim

Lexical Feature Extraction Method for Classification of Erroneous Online Customer Reviews Based on Pattern Matching 225
Maengsik Choi, Junsoo Shin, Harksoo Kim

Unified Concept Space and Mapping Discovery Algorithm for Heterogeneous Knowledge Systems 231
Lijun Zhu, Chen Shi, Jianfeng Guo

Author-Topic over Time (AToT): A Dynamic Users' Interest Model 239
Shuo Xu, Qingwei Shi, Xiaodong Qiao, Lijun Zhu, Hanmin Jung, Seungwoo Lee, Sung-Pil Choi

Scalable RDF Path Query Processing Based on Runtime Class Path Lookup Scheme 247
Sung-Jae Jung, Dong-min Seo, Seungwoo Lee, Hanmin Jung

Risk Aversion Parameter Estimation for First-Price Auction with Nonparametric Method 253
Xin An, Jiancheng Chen, Yuan Zhang

Diverse Heterogeneous Information Source-Based Researcher Evaluation Model for Research Performance Measurement 261
Jinhyung Kim, Myunggwon Hwang, Do-Heon Jeong, Sa-kwang Song, Jangwon Gim, Hanmin Jung, Shuo Xu, Lijun Zhu

Fast Big Textual Data Parsing in Distributed and Parallel Computing Environment 267
Jung-Ho Um, Chang-Hoo Jeong, Sung-Pil Choi, Seungwoo Lee, Hanmin Jung

K-Base: Platform to Build the Knowledge Base for an Intelligent Service 273
Sungho Shin, Jung-Ho Um, Sung-Pil Choi, Hanmin Jung, Shuo Xu, Lijun Zhu

A Novel Anomaly Detection System Based on HFR-MLR Method 279
Eunhye Kim, Sehun Kim

Knowledge Discovery and Integration: A Case Study of Housing Planning Support System 287
Junyoung Choi, Daesung Lee, Hanmin Jung

Data Intensive Computing and Applications

Performance Analysis of MapReduce-Based Distributed Systems for Iterative Data Processing Applications 293
Min Yoon, Hyeong-il Kim, Dong Hoon Choi, Heeseung Jo, Jae-woo Chang

A Semi-clustering Scheme for Large-Scale Graph Analysis on Hadoop 301
Seungtae Hong, Youngsung Shin, Dong Hoon Choi, Heeseung Jo, Jae-woo Chang

Multi-stream Parallel String Matching on Kepler Architecture 307
Nhat-Phuong Tran, Myungho Lee, Sugwon Hong, Dong Hoon Choi

Microscopic Bit-Level Wear-Leveling for NAND Flash Memory 315
Yong Song, Woomin Hwang, Ki-Woong Park, Kyu Ho Park

HASV: Hadoop-Based NGS Analyzer for Predicting Genomic Structure Variations 321
Gunhwan Ko, Jongcheol Yoon, Kyongseok Park

Multimedia Cloud Computing and Its Applications

Provisioning On-Demand HLA/RTI Simulation Environment on Cloud for Distributed-Parallel Computer Simulations 329
In-Yong Jung, Byong-John Han, Chang-Sung Jeong

Fast Shear Skew Warp Volume Rendering Using GPGPU for Cloud 3D Visualization 335
Chang-Woo Cho, Ki-Hyun Kim, Ki-Young Choi, Chang-Sung Jeong

A Vision-Based Robust Hovering Control System for UAV 341
Tyan Vladimir, Dongwoon Jeon, Doo-Hyun Kim

Finding Relationships between Human Affects and Colors Using SVD and pLSA 347
Umid Akhmedjanov, Eunjeong Ko, Yunhee Shin, Eun Yi Kim

Home Appliance Control and Monitoring System Model Based on Cloud Computing Technology 353
Yun Cui, Myoungjin Kim, Seung-woo Kum, Jong-jin Jung, Tae-Beom Lim, Hanku Lee, Okkyung Choi

Load Distribution Method for Ensuring QoS of Social Media Streaming Services in Cloud Environment 359
Seung Ho Han, Myoungjin Kim, Yun Cui, SeungHyun Seo, Yi Gu, Hanku Lee

A Robust Cloud-Based Service Architecture for Multimedia Streaming Using Hadoop 365
Myoungjin Kim, Seung Ho Han, Jong-jin Jung, Hanku Lee, Okkyung Choi

Video Image Based Hyper Live Spatial Data Construction 371
Yongwon Cho, Muwook Pyeon, Daesung Kim, Sujung Moon, Illwoong Jang

A Peer-to-Peer Based Job Distribution Model Using Dynamic Network Structure Transformation 377
Seungha Lee, Yangwoo Kim, Woongsup Kim

Distributed 2D Contents Stylization for Low-End Devices 385
Mingyu Lim, Yunjin Lee

Authority Delegation for Safe Social Media Services in Mobile NFC Environment 391
Jinsung Choi, Okkyung Choi, Yun Cui, Myoungjin Kim, Hanku Lee, Kangseok Kim, Hongjin Yeh

Mobile Computing

Introspection-Based Periodicity Awareness Model for Intermittently Connected Mobile Networks 397
Okan Turkes, Hans Scholten, Paul Havinga

Collaborative Recommendation of Mobile Apps: A Swarm Intelligence Method 405
Xiao Xia, Xiaodong Wang, Xingming Zhou, Tao Zhu

Enhanced Implementation of Max* Operator for Turbo Decoding 413
Dongpei Liu, Hengzhu Liu, Li Zhou

Ubiquitous Computing

A Context Description Language for Medical Information Systems 421
Kurt Englmeier, John Atkinson, Josiane Mothe, Fionn Murtagh, Javier Pereira

Eccentricity-Based Data Gathering and Diameter-Based Data Forwarding in 3D Wireless Sensor Networks 433
A.S.M. Sanwar Hosen, Gi-hwan Cho

Weighted Mining Frequent Itemsets Using FP-Tree Based on RFM for Personalized u-Commerce Recommendation System .. 441
Young Sung Cho, Song Chul Moon

The System of Stress Estimation for the Exposed Gas Pipeline Using the Wireless Tilt Sensor 451
Jeong Seok Oh, Hyo Jung Bng, Si-Hyung Lim

The Architecture Design of Semantic Based Open USN Service Platform Model 457
Hyungkyu Lee, Namje Park, Hyo-Chan Bang

Use-Cases and Service Modeling Analysis of Open Ubiquitous Sensor Network Platform in Semantic Environment 463
Taegyeong Kang, Namje Park, Hyungkyu Lee, Hyo-Chan Bang

A Neural Network Based Simple Weak Learner for Improving Generalization Ability for AdaBoost 469
Jongjin Won, Moonhyun Kim

Intelligent Computing

Serial Dictatorial Rule-Based Games for Camera Selection 475
Gowun Jeong, Yong-Ho Seo, Sang-Soo Yeo, Hyun S. Yang

Security Analysis on a Group Key Transfer Protocol Based on Secret Sharing 483
Mijin Kim, Namje Park, Dongho Won

Analysis of Cyber Attacks and Security Intelligence 489
Youngsoo Kim, Ikkyun Kim, Namje Park

Protection Profile for PoS (Point of Sale) System 495
Hyun-Jung Lee, Youngsook Lee, Dongho Won

Intelligent and Mobile Services

A Probabilistic Timing Constraint Modeling and Functional Validation Approach to Dynamic Service Composition for LBS 501
Weimin Li, Xiaohua Zhao, Jiulei Jiang, Xiaokang Zhou, Qun Jin

An Implementation of Augmented Reality and Location Awareness Services in Mobile Devices 509
Pei-Jung Lin, Sheng-Chang Chen, Yi-Hsung Li, Meng-Syue Wu, Shih-Yue Chen

Development of STEAM Education Program Centering on Non-traditional Energy 515
Yilip Kim, Jeongyeun Kim, Namje Park, Hyungkyu Lee

Scalable Key Management for Dynamic Group in Multi-cast Communication 521
Fikadu B. Degefa, Dongho Won

Result of Implementing STEAM Program and Analysis of Effectiveness for Smart Grid's Education 529
Jeongyeun Kim, Yilip Kim, Namje Park

Security Enhanced Unlinkable Authentication Scheme with Anonymity for Global Mobility Networks 535
Youngseok Chung, Seokjin Choi, Youngsook Lee, Dongho Won

3D Converged IT and Optical Communications

A Feature-Based Small Target Detection System 541
Jong-Ho Kim, Young-Su Park, Sang-Ho Ahn, Sang-Kyoon Kim

A Small Target Detection System Based on Morphology and Modified Gaussian Distance Function 549
Jong-Ho Kim, Jun-Jae Park, Sang-Ho Ahn, Sang-Kyoon Kim

Frontier Computing - Theory, Technologies and Applications

Using Hardware Acceleration to Improve the Security of Wi-Fi Client Devices 557
Jed Kao-Tung Chang, Chen Liu

An Anonymous Communication Scheme with Non-reputation for Vehicular Ad Hoc Networks 563
Ching-Hung Yeh, Meng-Yen Hsieh, Kuan-Ching Li

A Mobility Management Scheme for Internet of Things 569
Yuan-Kai Hsiao, Yen-Wen Lin

An Overlay Network Based on Arrangement Graph with Fault Tolerance 577
Ssu-Hsuan Lu, Kuan-Ching Li, Kuan-Chou Lai, Yeh-Ching Chung

Event Detection in Wireless Sensor Networks: Survey and Challenges 585
Aziz Nasridinov, Sun-Young Ihm, Young-Sik Jeong, Young-Ho Park

Accelerating Adaptive Forward Error Correction Using Graphics Processing Units 591
Md Shohidul Islam, Jong-Myon Kim

High-Performance Sound Engine of Guitar on Optimal Many-Core Processors 599
Myeongsu Kang, Cheol-Hong Kim, Jong-Myon Kim

Community Identification in Multiple Relationship Social Networks 609
Ting-An Hsieh, Kuan-Ching Li, Kuo-Chan Huang, Kuo-Hsun Hsu, Ching-Hsien Hsu, Kuan-Chou Lai

Intelligent System and Its Applications

An Improved ACO by Neighborhood Strategy for Color Image Segmentation 615
Shih-Pang Tseng, Ming-Chao Chiang, Chu-Sing Yang

A Novel Spiral Optimization for Clustering 621
Chun-Wei Tsai, Bo-Chi Huang, Ming-Chao Chiang

Recent Development of Metaheuristics for Clustering 629
Chun-Wei Tsai, Wei-Cheng Huang, Ming-Chao Chiang

The Originality of a Leader for Cooperative Learning 637
Po-Jen Chuang, Chu-Sing Yang

A Hybrid Ant-Bee Colony Optimization for Solving Traveling Salesman Problem with Competitive Agents 643
Abba Suganda Girsang, Chun-Wei Tsai, Chu-Sing Yang

Author Index 649

A Novel Ranking Technique Based on Page Queries*

Gwangbum Pyun and Unil Yun**

Department of Computer Engineering,
Sejong University, Republic of Korea
{pyungb,yunei}@sejong.ac.kr

Abstract. Keyword-based information retrieval finds webpages with queries composed of keywords to provide users with needed information. However, since the keywords are only a part of the necessary information, it may be hard to search intended results from the keyword-based methods. Furthermore, users should make efforts to select proper keywords many times in general because they cannot know which keyword is effective in obtaining meaningful information they really want. In this paper, we propose a novel algorithm, called PQ_Rank, which can find intended webpages more exactly than the existing keyword-based ones. To rank webpages more effectively, it considers not only keywords but also all of the words included in webpages, named page queries. Experimental results show that PQ_Rank outperforms PageRank, a famous algorithm used by Google, in terms of MAP, average recall, and NDCG.

Keywords: Information retrieval, Page query, Grouping webpages, Ranking technique.

1 Introduction

Most of recent information retrieval algorithms have been studied on the basis of keyword-based searches. There are well-known methods such as PageRank [4], RL_Rank [1], PDOM [3], etc., which help users obtain their intended webpages by applying various techniques. However, in these keyword-based approaches, users have to make efforts to select appropriate keywords several times in most cases since the keywords only represent a part of results they try to find. In this paper, to improve disadvantages of the existing methods, we propose a new ranking algorithm using webpages themselves as important queries in addition to the keywords, called PQ_Rank(Page Query_Rank), which is performed as follows. When certain keywords are inputted by a user, the algorithm first provides the user with webpages related to them. Then, the user selects one of the offered webpages, and thereafter, the algorithm compares relevance between the words of the selected one and those of other webpages. After that, PQ_Rank preferentially shows to the user webpages with higher

* This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2013005682 and 20080062611).
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_1,

relevance. PQ_Rank has two important features as follows. First, the algorithm can find more helpful webpages using the page query that is information more extended than the keyword query. Second, it can more efficiently perform comparison tasks between webpages by classifying them as a number of groups.

The remaining sections of this paper are as follows. In Section 2 and 3, related work is introduced, and the indexing method and ranking technique of the proposed algorithm are described respectively. In Section 4 and 5, we provide results of performance evaluation and conclude this paper.

2 Related Work

PageRank [4], as a ranking algorithm used by Google, assigns priorities per webpage after calculating the number of links between webpages to rank them. As another approach, RL_Rank [1] uses topics of webpages that users visit frequently as well as the keyword queries in order to rank webpages. In [7], its proposed algorithm stores user information to a database and compares similarity between queries and the information. RCW [6] selects a webpage first and then finds specific webpages related to the selected one. RCW sets weights of webpages based on *tf-idf* [5], so it has the most effect on ranking scores of RCW. Meanwhile, the proposed algorithm, PQ_Rank does not depend on the *tf-idf* and uses a new-type ranking technique using the page query based on webpage groups classified by the indexer of the algorithm.

3 A Ranking Technique Based on Page Queries

3.1 Indexing Web Pages

Before calculating ranking scores with queries, we first collect webpages, classify them as several groups, and then index them. The indexer of the proposed algorithm has structures in the form of a group, where each group has a linear list storing webpages and a comparison list having all words in the group. Fig. 1 shows how the indexer operates, and its detailed steps are as follows. First, the words of a newly collected webpage are compared to the ones of comparison lists stored for each group.

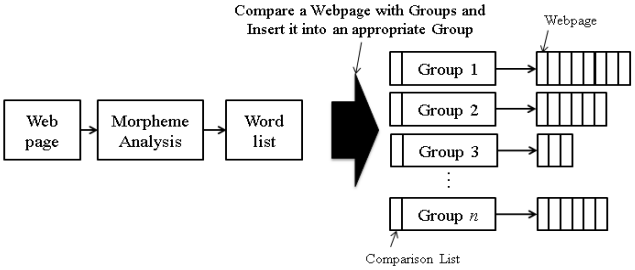


Fig. 1. Execution processes of the indexer

Thereafter, the new webpage is inserted into an appropriate group which has the most words matched with those of the new webpage, where certain words which do not exist in the comparison list of the selected group are updated to the list. To insert the new webpage into the group properly, the algorithm first searches a certain webpage having the highest congruence within the current group, comparing the words of the new webpage with those of the webpages included in the group. The new one is inserted in front of the searched one. If there is no appropriate group, a new group is generated and the new webpage and its word list are inputted in the new group. Generation of the new group is allowed when the congruence between the new webpage and each group is less than N%(N is set by PQ_Rank). The congruence means how many the words of the newly inserted webpage is matched with those of the comparison list in any group. After the insertion phases terminate with respect to all of the collected webpages, the algorithm creates an inverted file and group-related files. The inverted file stores relative frequencies of words, URL addresses, and webpage positions for each group. The positions have group and list information of the webpages, i.e., indicate where the webpages are located. In the group-related files, for each group, all of the groups except for it are sorted in descending order of their congruence and the sequences are stored. In other words, the algorithm makes this file per group, where the file stores other groups' names and sequences according to their congruence.

3.2 Ranking Technique

PQ_Rank shows webpages sorted in descending order of keyword frequency at first. Then, among the shown pages, a user selects one which is likely to contain necessary information. After that, PQ_Rank calculates ranking scores so that webpages most related to the selected one can be ranked as the top positions, where we have only to consider webpages with the first inputted keyword as target pages that compute ranking scores. The scores are calculated as follows.

$$Ranking\ Score = \begin{cases} \frac{size\ of\ G_{P,Q}}{|P_{location} - Q_{location}|}, if\ P\ and\ Q\ are\ in\ the\ same\ group \\ \frac{1}{con_{G_P,G_Q} - Q_{tf} + 1}, otherwise \end{cases}$$

Let P be a webpage selected by the user and Q be a webpage to calculate its ranking score. If both P and Q belong to the same group, G, The ranking score is obtained by dividing the size of G including P and Q, $size\ of\ G_{P,Q}$ by the distance of them, $|P_{location} - Q_{location}|$, where the value is always more than 1. If they are not contained in the same group, we compute the score as a different manner in order to consider congruence between groups. Let con_{G_P,G_Q} be congruence between the groups including P and Q respectively, denoted as G_P and G_Q. Then, it has an integer value from 1(as the two groups are more similar) to the total number of groups(as they are less relevant). In addition, we use Q's term frequency, Q_{tf}($0 \leq Q_{tf} \leq 1$) to distinguish importance of webpages which are not included in G_P but

are contained in the same group to each other. Therefore, webpages with higher *tf*(term frequency) are likely to have relatively better ranking scores. To differentiate these two cases, we add 1 to the denominator of the second formula. Thereby, the first case always returns more than 1 value while the second one has less than 1 and more than 0. That is, specific webpages belonging to the same group as that of the selected page have relatively high ranking scores.

4 Performance Evaluation

In this section, we conduct performance evaluation regarding the proposed algorithm, PQ_Rank and Google. Our algorithm was written in C++ and, we used CLucene [8] to conduct morpheme analysis. The standard for generating new groups, N was set as 10%. Data used for the experiments are newspaper articles from 1/1/2011 to 12/31/2011, which are collected at http://www.washingtonpost.com. The number of them is 42,150 and their categories are composed of politics, business, life, style, entertainment, sport, region, and world. Performance evaluation is conducted as follows. If random keywords are inputted, PQ_Rank first shows webpages according to their keyword frequencies, where one webpage is selected and the algorithm calculates ranking scores as mentioned in the section 3. Then, we evaluate the results. In Google, we measure its performance using the same keywords as those of PQ_Rank and important keywords of the selected webpages. The first measure, MAP(Mean Average Precision) [3] is to compute a mean value of average precision regarding all queries. Recall is a ratio of correct answer pages successfully searched from whole ones. Average recall is a mean recall value considering all of the queries. Fig. 2 shows results of MAP and average recall, where the test was performed with respect to the top-50 webpages and 12 random queries. Each query is composed of two elements(a keyword / an important word of the selected webpage), so all of the used queries are { (cup / golf), (Korea / nuclear), (apple / food), (sea / travel), (skin / health), (stone / football), (open / book), (war / obituaries), (French / economy), (windows / garden), (island / travel), and (wife / style)}. From the results of the experiment, we can observe that PQ_Rank has outstanding performance that that of Google in terms of the MAP and average recall measures. In Fig. 3, NDCG [3] is evaluated, where it is more likely to have higher values when correct answer pages are located in the top lank positions.

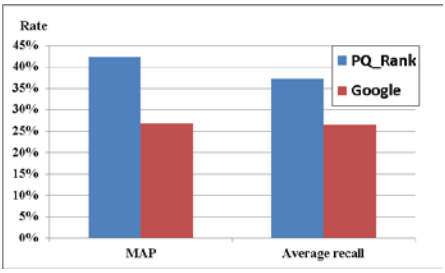


Fig. 2. MAP and average recall test

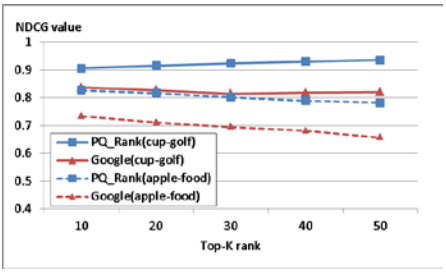


Fig. 3. NDCG test

NDCG is computed using DCG(Discounted Cumulative Gain) and IDCG(Ideal Discounted Cumulative Gain) as the following formula.

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i} \quad , \; IDCG_k = srel_1 + \sum_{i=2}^{k} \frac{srel_i}{log_2 i}$$

rel_i is a relevance value and set as 3(when the current webpage and the selected one have similar contents and the same category), 2(when they have similar contents but different categories), 1(when they have the same category but different contents), or 0(when they are entirely different). srel_i is to sort rel_i in score descending order. Thus, NDCG can calculate a gap of results between the ideal case and real ranking algorithms. Fig. 3 presents NDCG results for the queries (cup / golf) and (apple / food). As in the previous test, PQ_Rank guarantees better performance than that of Google in every case.

5 Conclusion

In this paper, we proposed a novel ranking algorithm, PQ_Rank which could perform more exact webpage searches by using webpages as queries. By using the newly proposed technique, page query, PQ_Rank could provide users with more meaningful search results compared to Google. These advantages were proved through the experimental results shown in this paper. In future work, we can consider applying PQ_Rank to various areas such as retrieving twitter, blog, and so on. These applications are expected to make a great contribution to the information retrieval area.

References

1. Derhami, V., Khodadadian, E., Ghasemzadeh, M., Bidoki, A.M.: Applying reinforcement learning for web pages ranking algorithms. Applied Soft Computing 13(4), 1686–1692 (2013)
2. Ermelinda, O., Massimo, R.: Towards a Spatial Instance Learning Method for Deep Web Pages. In: Industrial Conference on Data Mining, pp. 270–285 (December 2011)
3. Geng, B., Yang, L., Xu, C., Hua, X.S.: Ranking Model Adaptation for Domain-Specific Search. IEEE Transactions on Knowledge and Data Engineering 24(4), 745–758 (2012)
4. Ishii, H., Tempo, R., Bai, E.: A Web Aggregation Approach for Distributed Randomized PageRank Algorithms. IEEE Transactions on Automatic Control 57(11), 2703–2717 (2012)
5. Metzler, D.: Generalized Inverse Document Frequency. In: Conference on Information and Knowledge Management, pp. 399–408 (October 2008)
6. Pyun, G., Yun, U.: Ranking Techniques for Finding Correlated Webpages. In: International Conference on IT Convergence and Security, pp. 1085–1095 (December 2012)
7. Telang, A., Li, C., Chakravarthy, S.: One Size Does Not Fit All: Toward User- and Query-Dependent Ranking for Web Databases. IEEE Transactions on Knowledge and Data Engineering 24(9), 1671–1685 (2012)
8. CLucene Project web page, `http://clucene.sourceforge.net/`

Ranking Book Reviews Based on User Discussion*

Heungmo Ryang and Unil Yun**

Department of Computer Engineering,
Sejong University, Republic of Korea
{riangs,yunei}@sejong.ac.kr

Abstract. Ranking algorithm is one of the most important issues in information retrieval researches. It can be divided into two types according to purpose, general and specific purpose algorithms. Although the general purpose algorithms can be effective for retrieving relevant documents on the internet, it is difficult to find meaningful information of specific targets such as blogs or twitter since the algorithms do not consider unique characteristics of the targets. Recently, ranking algorithms have been proposed for searching useful book reviews by reflecting unique features of book reviews. In this paper, we propose a novel algorithm, RUD (Ranking based on User Discussion), for ranking book reviews to improve performance based on user discussion. For performance evaluation, we conduct precision and recall tests. The experimental results show that the proposed algorithm outperforms previous algorithms.

Keywords: Book review, Information retrieval, Ranking, User discussion.

1 Introduction

IR (Information Retrieval) systems are used to search relevant information from a data collection. In the system, information retrieval begins when a query is given from a user, and the query does not generally identify a single document in the collection. It means that some documents can be matched to the query with relevance degrees (or ranking scores). Ranking algorithm is adopted and employed to compute the ranking scores of keywords in documents of the collection. Hence, the ranking algorithm is one of the most important issues in information retrieval researches. On the other hand, many ranking algorithms [2, 5, 6] were proposed to retrieve meaningful information on the internet with the increasing amount of documents. These algorithms can be divided into two types according to purpose, general and specific purpose algorithms. It is difficult for IR systems with the general purpose algorithms [2] to find useful information of specific targets such as blogs or twitter. The reason is that the algorithms do not consider unique characteristics of the targets. Recently, ranking algorithms for specific purpose, searching meaningful book reviews, by

* This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_2,

reflecting features of the reviews were proposed [3, 4, 7]. In this paper, we propose a novel algorithm for ranking book reviews to improve performance of the previous algorithms. The remainder of this paper is organized as follows. In Section 2, we describe some influential related works. Section 3 illustrates our ranking algorithm in detail. In Section 4, we present and analyze experimental results for performance evaluation. Finally, our contributions are summarized in Section 6.

2 Related Work

LengthRank and ReplyRank were first proposed [4] for ranking book reviews. The algorithms are based on the fact that the probability of including the more additional information about books is higher when the length of contents is longer or the number of replies to reviews is larger. The experimental results in the study [4] showed that LengthRank and ReplyRank can find meaningful book reviews more effectively than a traditional keyword-based ranking technique, TF-IDF (Term Frequency-Inverse Document Frequency) [1]. RLRank [7] was afterward proposed to apply opinions about books, contents, and additional opinions about the reviews, replies, to the ranking together. The performance of the algorithm is better when the adoption rate of the number of replies is higher than that of the length of contents. In this paper, therefore, we propose a novel algorithm to reinforce performance of ranking reviews with replies based on user discussion. Meanwhile, ranking algorithms with ERQ (Estimated Reviewer Quality), LRERQ and RRERQ, were proposed [3] for considering reviewer quality as well as the amount of additional information.

3 The Proposed Ranking Method: RUD

In this Section, we propose a novel ranking algorithm based on user discussion, RUD (Ranking based on User Discussion), for searching meaningful book reviews effectively. Previous algorithms [3, 4, 7] focus on the amount of information in reviews, the number of replies, the length of contents, and the number of reviews written by each user. The proposed method, RUD considers an aspect of user discussion through replies as well as the number of reviews.

3.1 Ranking Book Reviews Based on User Discussion

A book review is an opinion of a reviewer about a book, and a reply of other user is an opinion about the review or a reply of another user. That is, users including the reviewer can discuss through replies. In addition, the result of the discussion can lead to refine the review to be more useful, especially when the reviewer participates in the discussion more actively.

Definition 1. *Involving rate* of a reviewer u_R in a review r refers to the rate of the number of replies of the reviewer rep_R by the total number of replies rep_T in r. It is denoted as $ir(u_R, r)$ and defined as $1 - \left|0.5 - \frac{rep_R}{rep_T}\right| \ (rep_T > 0)$. If there is no replies

in r, then $ir(u_R, r)$ becomes zero. For example, if there are 10 replies to a review and two among the replies are written by a reviewer, then the involving rate of the reviewer is $1 - \left|0.5 - \frac{2}{10}\right| = 1 - 0.3 = 0.7$.

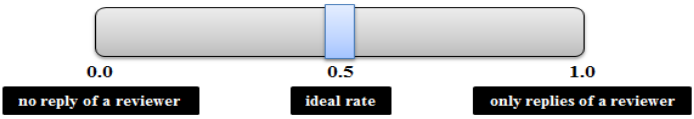


Fig. 1. The rate of reviewer's replies by the whole replies in a review

In the equation of the involving rate, $\frac{rep_R}{rep_T}$ indicates the rate of reviewer's replies by the whole replies in a review, and Fig. 1 shows the rate. In the figure, 0.5 refers to the ideal discussion which means that the reviewer comments to each reply of other user. In this case, the involving rate has a maximum value, 1. Otherwise, if there is no reply or are only replies of the reviewer, the value becomes 0.5, a minimum involving rate. Meanwhile, RUD is a keyword-based ranking, and thus we reflect the importance of keywords with TF-IDF [1]. The TF-IDF value of a keyword *key* in a review r, $tfIdf(key, r)$, is calculated by the multiplication *tf* by *idf*, where *tf* is the frequency of key in r and *idf* is a measure of whether the keyword is common or rare across a dataset. The *idf* value is computed by $log_2^N - log_2^{d_k} + 1$, where N is the total number of reviews in the dataset and d_k is the number of reviews containing the keyword.

Definition 2. *RUD* of a keyword *key* in a review r refers to a ranking score of *key* in r. It is denoted ad $RUD(key, r)$ and defined as $tfIdf(key, r) + tfIdf(key, r) \times ir(u_R, r)$.

INPUT	***BRD*: a given book review dataset, *SWL*: a stopword list**
OUTPUT	***RUDL*: a list of RUD ranking scores**

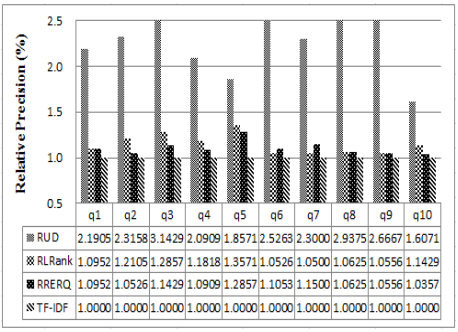
Let $IRL(rid, ir)$ be a involving rate list, where rid is an id of a review and ir is the involving rate
Let $FRL(rid, key, fr)$ be a list, where key is a keyword and fr is the frequency of key in a review
01: $N \leftarrow 0$
02: **For each** review r with a review id rid, a reviewer u_R, and reply list RPL in BRD
03: **Extract** keywords from r
04: **For each** keyword key in the extracted keywords
05: **If** key is in SWL then **continue**
06: **Count** the frequency fr of key in r and **Add** (rid, key, fr) to FRL
07: **Count** the total number of replies rep_T in RPL
08: **If** rep_T is zero then **Add** $(rid, 0)$ to IRL
09: **Else**
10: **Count** the number of u_R's replies rep_R in RPL
11: $ir \leftarrow 1 - |0.5 - (rep_R / rep_T)|$
12: **Add** (rid, ir) to IRL
13: $N \leftarrow N + 1$
14: **For each** keyword key with rid and fr in FRL
15: **Count** the number of reviews including key d_k in FRL
16: **Calculate** $tfIdf$ of key with N and d_k
17: $ir \leftarrow$ the involving rate in IRL such as (rid, ir)
18: $RUD \leftarrow tfIdf + tfIdf \times ir$
19: **Add** (rid, key, RUD) to $RUDL$

Fig. 2. The Proposed Ranking Algorithm of RUD

The proposed framework consists of four steps. In the first step, crawler (or web robot) collects book reviews on the internet. In the second step, analyzer extracts information such as contents, keywords, and replies from the collected data. In the third step, the ranking algorithm, RUD, is applied to compute ranking scores. Finally, indexer constructs an inverted file system for keyword-based searching. Fig. 2 shows a ranking algorithm of RUD for searching meaningful book reviews. First, for each review in the collected dataset, the frequency of each keyword is counted except stopwords and the involving rate of the review is calculated (lines 2 to 12). Then, TF-IDF values are computed with the counted frequency (lines 15 to 16). Finally, ranking scores of keywords are obtained with the TF-IDF values and the involving rates (lines 17 to 19). After construction of an inverted file system with the ranking scores, searching process is performed when a query is given from a user.

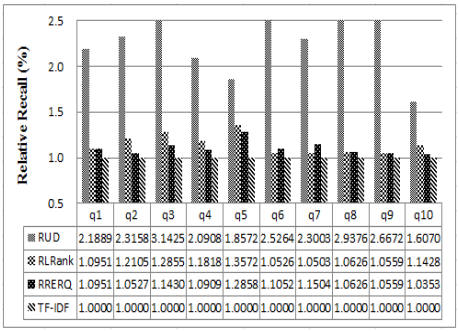
4 Performance Evaluation

In this Section, we evaluate performance of the proposed algorithm, RUD, with the previous algorithms, RLRank [7], RRERQ [3], TF-IDF [1]. The experiments were conducted on a 3.3 GHz Intel Processor with 8 GB memory, and run with the Windows 7 operating system. All algorithms used in this Section were written in C++ language. We have collected 0.11 million book reviews from one of the most famous online bookstores in the world, Amazon (http://www.amazon.com), and an online book review site, GoodReads (http://www.goodreads.com). Moreover, we measure the usefulness of reviews by the number of recommendations to the reviews. If a review has no less the number of recommendations than a minimum threshold value, then the review is useful in a dataset. To determine the threshold, we have analyzed our collected dataset. Then, the average value is 4.303, and we use the value as a minimum threshold. We employ 10 queries: history (q1), economy (q2), families (q3), photography (q4), biographies (q5), education (q6), humor (q7), entertainment (q8), relationships (q9), fantasy (q10). They are selected from category names of Amazon for the performance evaluation since the compared algorithms are keyword-based.



	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
RUD	2.1905	2.3158	3.1429	2.0909	1.8571	2.5263	2.3000	2.9375	2.6667	1.6071
RLRank	1.0952	1.2105	1.2857	1.1818	1.3571	1.0526	1.0500	1.0625	1.0556	1.1429
RRERQ	1.0952	1.0526	1.1429	1.0909	1.2857	1.1053	1.1500	1.0625	1.0556	1.0357
TF-IDF	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(a)



	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
RUD	2.1889	2.3158	3.1425	2.0908	1.8572	2.5264	2.3003	2.9376	2.6672	1.6070
RLRank	1.0951	1.2105	1.2855	1.1818	1.3572	1.0526	1.0503	1.0626	1.0559	1.1428
RRERQ	1.0951	1.0527	1.1430	1.0909	1.2858	1.1052	1.1504	1.0626	1.0559	1.0353
TF-IDF	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(b)

Fig. 3. Performance evaluation of precision and recall

We first compare performance of RUD with RLRank, RRERQ, and TF-IDF by precision test. Precision indicates the rate of the number of retrieved relevant reviews by that of searched *K* reviews, and *K* is set to 50 in the experiments of this Section. Fig. 3a shows result of the precision test, and values in the figure are the relative rate of precisions of the compared algorithms by precision of TF-IDF. The precision of RUD is the best, followed by both RLRank and RRERQ, and TF-IDF is the worst. On the other hand, the performance of RLRank is better than that of RRERQ (q2, q3, q4, q5, and q10), worse (q6), and the same (q1, q7, q8, and q9). Hence, we cannot say which one of them is more effective for ranking book reviews.

Finally, we conduct recall test of the compared algorithms. Recall refers to the rate of the number of retrieved relevant reviews by the total number of relevant reviews with respect to a query. Fig. 3b is result of the recall test, and *K* is assigned to 50 as the precision test. In the figure, we can observe that the performance of TF-IDF is the worst, followed by both RLRank and RRERQ, and our ranking algorithm, RUD, is the best. Furthermore, we can know that the relative performances of the compared algorithms on precision and recall are similar. From the experimental results, we can learn that active user discussion is a significant factor in considering reviews as meaningful to users.

5 Conclusions

In this paper, we proposed a novel ranking algorithm, RUD (Ranking based on User Discussion), for searching meaningful book reviews. Furthermore, we conducted experiments of precision and recall for performance evaluation of the proposed algorithm. The experimental results showed that our algorithm, RUD, outperformed the previous ranking algorithms, RLRank, RRERQ, and TF-IDF, in terms of precision and recall. Moreover, we analyzed that the factor based on user discussion is effective on retrieving useful book reviews. As a future work, this study can be applicable to recommendation systems for meaningful book reviews to users.

References

1. Aizawa, A.N.: An Information-theoretic Perspective of TF-idf Measures. Journal Information Processing and Management 39(1), 45–65 (2003)
2. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab (1999)
3. Ryang, H., Yun, U.: Book Review Retrieval Techniques for Adopting Estimated Reviewer Quality. In: Lee, G., Howard, D., Kang, J.J., Ślęzak, D. (eds.) ICHIT 2012. LNCS, vol. 7425, pp. 550–557. Springer, Heidelberg (2012)
4. Ryang, H., Yun, U.: Effective Ranking Techniques for Book Review Retrieval Based on the Structural Feature. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. LNCS, vol. 6935, pp. 360–367. Springer, Heidelberg (2011)
5. Sarma, A.D., Sarma, A.D., Gollapudi, S., Panigrahy, R.: Ranking mechanisms in twitter-like forums. In: WSDM, New York, pp. 21–30 (2010)
6. Weerkamp, W., Rijke, M.: Credibility-inspired ranking for blog post retrieval. Information Retrieval 15(3-4), 243–277 (2012)
7. Yun, U., Ryang, H., Pyun, G., Lee, G.: Efficient Opinion Article Retrieval System. In: Lee, G., Howard, D., Kang, J.J., Ślęzak, D. (eds.) ICHIT 2012. LNCS, vol. 7425, pp. 566–573. Springer, Heidelberg (2012)

The Blog Ranking Algorithm Using Analysis of Both Blog Influence and Characteristics of Blog Posts*

Jiwon Kim and Unil Yun**

Department of Computer Engineering,
Sejong University
{jwonkim,yunei}@sejong.ac.kr

Abstract. In recent years, while amounts of the information in the blogosphere increase rapidly, the problems of information quality have come up. Discovering a good quality data is the important issue in blog space with overflowed information. In this paper, we present WCT algorithm for efficient blog ranking. This method performs a ranking process using both interconnection of blogs and structural weights for content in blog. In the performance analysis, we discuss the comparison between our algorithm and the previous algorithm for blog ranking. The result shows that our proposal has the high performance than other blog retrieval method.

Keywords: Blog ranking, Information retrieval, Blog structure.

1 Introduction

A Blog is an informational website and consists of discrete entries, known as posts, authored by a blogger who manages a blog, and the blogosphere[8], which is the set of all blogs, has grown to the enormous size according to rapidly development of web services. However, this growth provokes the uncertainty of information due to increasing of the incorrect information such as spam. To solve the problem in the blogosphere, many algorithms[1, 2, 4, 5] to estimate the quality of the posts have been devised in the field of information retrieval. In this paper, we propose an efficient algorithm for blog search, WCT(Weighted Comment and Trackback) algorithm. The WCT algorithm uses many properties of the blog posts. Especially, this approach focuses on the comments and trackbacks. Hence, we conduct the performance evaluation between the WCT algorithm and previous algorithm, and this result shows that the proposed method outperforms other method.

This paper is organized as follows. Section 2 briefly introduces the various measures of information retrieval. In Section 3, we describe the suggested algorithm.

* This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_3,

Section 4 presents performance evaluation including experimental environments. Finally, Section 5 addresses the conclusion of this paper.

2 Related Work

To more efficient and reliable search, many researchers have studied the field of information retrieval[1, 2, 4, 5, 7]. In particular, there are a variety of algorithms that are utilized in the blog ranking. These algorithms are classified as two categories of content-based algorithms[2] and connectivity-based algorithms[1, 4, 5]. First, content-based algorithms decide similarity between the specific keywords and documents when the keywords are given by users. The content-based algorithms include tf-idf[2] strategy. Second, connectivity algorithms, known as link-based, use the connection among blogs(e.g. comments, trackbacks, links) in the blogosphere. There is PageRank algorithm which is the most famous algorithm of these methods. The PageRank algorithm estimates the importance of web page by the number of links received from other web pages. The HITS(Hyperlink-Induced Topic Search)[5] method calculates the ranking of blogs by consideration of the number of input-output links. Recently, the various models[7] for measuring the impact of blogs have been proposed.

3 The WCT Algorithm

In this Section, we suggest WCT(Weighted Comment and Trackback) algorithm using interconnection among blogs and the structural characteristic of blog posts to the efficient blog ranking. Specially, the WCT algorithm concentrates upon the comments and trackbacks since the blogger's behavior writing the comments and trackbacks means to communicate between bloggers. That is, the more active blogging(i.e. blogger uses blog) makes the higher quality of blog posts. The WCT algorithm is computed as follows:

$$WCT(B,P,K) = NCS(B,P) + BSS(P,K) \quad (1)$$

In Eq. (1), the input parameters are keyword *K*, post *P* containing the keyword *K*, and blog *B* including post *P*, respectively. *NCS(Network Connectivity Score)* measures the influence of blogs in blogosphere, and *BSS(Blog Structural Score)* estimates the score of weighted *K* using the analysis of the structural characteristics of blogs.

3.1 Estimating the Network Connectivity Score

In this approach, we derive the influence of each blog by analyzing the link relationships of the connections between blogs and using the SPEAR(Spamming-resistant expertise analysis and ranking)[9] algorithm. The *NCS(B, P)* is defined as

$$NCS(B,P) = log_2\{1 + \alpha \times CMT(P) + \beta \times TRBK(P)\} + SPEAR(B) \quad (2)$$

In Eq. (2), α and β are parameters with regard to weights for each blog structure about comment and trackback, respectively. The *CMT(P)* counts the number of comments of a post P. The number of comments indicates the popularity of blog post because having many comments in the blog post is the measure of high user's reactivity. After that, the *TRBK(P)* analyzes connection of trackbacks. The trackback functions as that automatically generates a reverse link and creates a link between the blog posts. Furthermore, we add the outcome of *SPEAR(B)* to result. The SPEAR algorithm analyzes the connectivity with other pages and simply quantifies the connections. At this time, target of analysis includes that blogroll(i.e. blogger's list of hyperlinks to other blogs of recommend) and many links in the posts.

3.2 Estimating Blog Structure Score

The blog has some structural characteristics which could not find in other types of web pages. In our proposal, we use these properties to find the topic of blog posts. The blog post is typically composed: title, body, tags(set of keywords chosen by blogger), date, URL, comments, trackbacks, and RSS feed. Whenever WCT algorithm extracts the terms in the blog post, we assign a weight according to the structure of the extracted term. In addition, we also use the weights for the emphasized texts(e.g. bold, italic, underline) and hyperlink texts. In other words, the extracted term from specific structure that is exposed to topics has more weight. *BSS(P, K)* use the following formula:

$$BSS(P,K) = PIDF(K) \times \sum_{i=0}^{n} \{ W_i(P,K) \times TF_i(P,K) \} \quad (3)$$

In Eq. (3), the parameter *n* is frequency of keyword *K* found in the post *P*. If keyword *K* are faced *n* times in the post, this metric computes all of the weighted term frequency. After that, the result value is multiplied the outcome of *PIDF(K)*. The *BSS(P, K)* computes not only the term frequency in blog post but also the term frequency in the entire posts due to the noise such as spams in the blogosphere. *PIDF(K)* is denoted as follows:

$$PIDF(K) = log_2 \frac{the\ Number\ of\ Total\ Blog\ Posts}{|\, d_K - 17\,| + 1} \quad (4)$$

In Eq. (4), d_K is the number of posts including keyword *K*. PIDF[6] method based on the idf metric considers an additional element(number 17, optimized value). Thus, terms possessed with middle frequency have more large weight than others. When we index the blog posts, this approach facilitates the high quality result excluded spam or unnecessary things such as preposition.

4 Performance Analysis

In this section, we show the performance comparison between blog ranking algorithms. We explain the experimental environment, and then we present the results of evaluation for set of real data.

4.1 Experimental Environment

The Experiments were performed on Windows 7 OS, with Intel quad-core processor(i5-3570 3.4GHz), 8GB memory, and MySQL 5.6. In addition, the entire algorithms were written by Python language. For performance evaluation, we have used the real blog data. The dataset crawled from the famous blog search site, Technorati (http://www.technorati.com). This is described in the table as below:

Table 1. Statistics of entire indexed content used in the paper

Crawling Period	*04/2013*
The number of blogs	2,157
The number of blog posts	43,781
The average of blog posts per blog	19.286
The number of words	239,720

The dataset has collected with 43,781 posts in 2,157 blogs and the two types of blogs of Wordpress(http://wordpress.com/) and Blogspot(http://www.blogger.com/).

4.2 Experimental Results

In order to evaluate our proposed approach, we performed the NDCG(Normalized Discounted Cumulative Gain)[3] test. In the field of information retrieval, the NDCG measures effectiveness of the information retrieval algorithms, and uses the relevance between query and result set. In this paper, the length of body of blog post was used as factor of NDCG because the length of body is criterion representing amount of information. Moreover, the WCT algorithm is compared with Trackback-Rank Algorithm[4]. Trackback-Rank Algorithm only uses trackback connectivity among blogs to evaluate authority of blog and score of post's reactivity in order to rank blog posts.

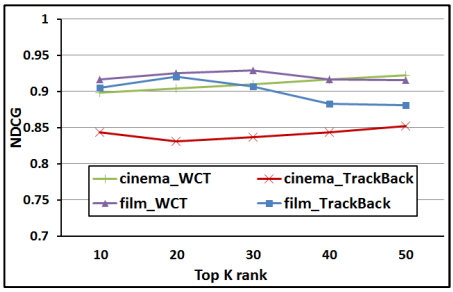


Fig. 1. NDCG at top K for *cinema* and *film*

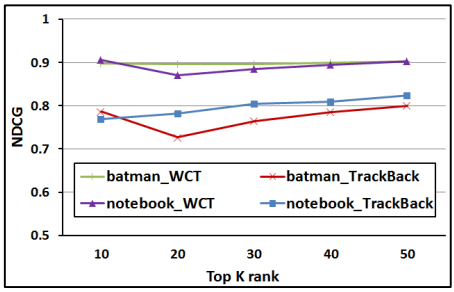


Fig. 2. NDCG at top K for *batman* and *notebook*

Fig. 1 and Fig. 2 represent the NDCG at top K scores for each query. Fig. 1 is NDCG scores given the queries, *cinema* and *film*. On the whole, all NDCG scores of WCT are high compared with the Trackback-Rank algorithm, although the score is very similar when query is *film* at top 20 rank. Fig. 2 is the result of the query *batman* and *notebook*. As shown in the graph, WCT algorithm has the always highest scores in the graph. Consequently, this experimental result using NDCG metric presented that WCT algorithm is better performance than Trackback-Rank algorithm.

5 Conclusion

In this paper, we proposed an algorithm, which is called the WCT(Weighted Comments and trackbacks), that efficiently performs a ranking for the blog retrieval in order to solve the uncertainty of information in blogosphere. Therefore, WCT rank algorithm uses the linkage patterns such as blogrolls, trackbacks, and hyperlinks among blog posts to evaluate the importance of blogs. In addition, the WCT algorithm was adapted the weights to words based on the blog structure when the words are extracted in the posts, and then the pivot-idf method conducted to normalize the result. In experimental evaluation, the comparison of our proposal and other blog ranking algorithm, called Trackback-Rank, was conducted by NDCG metric. As a result, WCT algorithm presented higher performance than previous blog ranking algorithm. The study of the future, we will research the extraction of topic of blog posts using the clustering techniques of data mining and then this technique utilizes the blog retrieval system.

References

1. Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine. In: Proceedings of 7th International World Wide Web Conference, Computer Networks and ISDN Systems, vol. 30(1-7), pp. 107–117 (1998)
2. Fautsch, C., Savoy, J.: Adapting the Tf-Idf Vector-space Model to Domain specific Information Retrieval. In: SAC 2010, pp. 1708–1712 (2010)
3. Jarvelin, K., Kekalainen, J.: IR evaluation methods for retrieving highly relevant documents. In: SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48 (2000)
4. Kim, l.H., Yoon, T.B., Kim, K.S., Lee, J.-H.: The Trackback-Rank Algorithm for the Blog Search. In: IEEE International Multi-topic Conference 2008, pp. 454–459 (2008)
5. Kleinberg, J.M.: Authoritative sources in hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
6. Lee, J.: A Study on the Pivoted Inverse Document Frequency Weighting Method. Journal of the Korea Society for Information Management 20(4), 233–248 (2003)
7. Momma, M., Chi, Y., Lin, Y., Zhu, S., Yang, T.: Influence Analysis in the Blogosphere. The Computing Research Repository, CoRR 2012, abs/1212.5863 (2012)
8. Santos, R.L.T., Macdonald, C., McCreadie, R.M.C., Ounis, I., Soboroff, I.: Information Retrieval on the Blogosphere. Foundations and Trends in Information Retrieval 6(1), 1–125 (2012)
9. Yeung, C.M.A., Noll, M.G., Gibbins, N., Meinel, C., Shadbolt, N.: SPEAR: Spamming-Resistant Expertise Analysis and Ranking in Collaborative Tagging Systems. Computational Intelligence 27(3), 458–488 (2011)

Frequent Graph Mining Based on Multiple Minimum Support Constraints*

Gangin Lee and Unil Yun**

Department of Computer Engineering, Sejong University, Republic of Korea
{abcnarak,yunei}@sejong.ac.kr

Abstract. Frequent graph mining allows us to find useful frequent sub-graph patterns from large and complicated graph databases. In a lot of real world applications, graph patterns with relatively low supports can be used as meaningful information. However, previous methods based on a single minimum support threshold have trouble finding them. That is, they cause "*rare item problem*", which means that useful sub-graphs with low supports cannot be extracted when a minimum support threshold is high, while an enormous number of patterns have to be mined to obtain these useful ones when the value is low. To overcome this problem, we propose a novel algorithm, FGM-MMS (Frequent Graph Mining based on Multiple Minimum Support constraints). After that, we demonstrate that the suggested algorithm outperforms a state-of-the-art graph mining algorithm through comprehensive performance experiments.

Keywords: Frequent graph mining, Multiple minimum supports, Sub graph.

1 Introduction

Graphs are useful data structure which can express almost all data from the real world effectively. Therefore, many studies for mining these graph data[3, 5, 6] have been conducted. Most graph databases obtained from the real world include not only high frequent sub-graph patterns but also meaningful ones with relatively low frequency(or support). However, since previous methods consider a single minimum support threshold, there occurs the following problem, i.e., *rare item problem*[4]. When the minimum support threshold is high, they cannot find meaningful sub-graph patterns with small supports while when it is low, they have to mine an enormous number of patterns to find the useful ones. In this paper, to solve the problem, we propose a novel graph mining algorithm applying multiple minimum support constraints, called FGM-MMS(Frequent Graph Mining based on Multiple Minimum Support constraints). The algorithm allows us to obtain useful sub-graphs without lowering the minimum support threshold more than needs. Experimental results provided in this

* This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_4,

paper show that the suggested algorithm has outstanding performance more than that of a state-of-the-art algorithm.

The remainder of this paper is organized as follows. In Section 2 and 3, we introduce related work and details of the proposed algorithm, FGM-MMS. In Section 4 and 5, we analyze results of our comprehensive experiments and conclude this paper.

2 Related Work

To solve the *rare item problem* from databases composed of itemsets, various frequent pattern mining approaches[1, 2, 4] have been proposed. MSApriori[4] is a level-wise method for solving this problem. A FP-growth-based algorithm, CFP-growth[1] solves the same issue more efficiently. CFP-growth++[2] is its advanced version. They conduct mining works considering different minimum item supports for each item, not a single value. Therefore, they can easily mine specific itemsets with low supports but important characteristics. However, since they are limited to mining itemset-based simple databases, they do not deal with complicated graph databases.

Among many fundamental graph mining algorithms, Gaston[5, 6], which is a state-of-the-art algorithm, performs the most efficient mining operations. The algorithm extracts frequent sub-graph patterns classifying its mining procedure into the three phases: *path*, *free tree*, and *cyclic graph*. In addition, it uses *embedding list* to reduce runtime needed for mining patterns. Our algorithm is compared with the Gaston for fair performance comparison. To our knowledge, FGM-MMS is the first graph mining algorithm which can solve the *rare item problem*. By applying the multiple minimum support constraints, it can extract meaningful sub-graph patterns considering supports for each element of graphs.

3 Mining Frequent Sub-graph Patterns Based on Multiple Minimum Support Constraints

3.1 Applying Multiple Minimum Support Constraints on the Graph Mining

In contrast to itemsets composed of only items, graph patterns consist of a number of elements such as vertices and edges. Therefore, the graph mining approach should consider all the elements. Furthermore, in this multiple minimum support framework, we have to consider different minimum element support values for each element.

Definition 1. (*MES*(*Minimum Element Support*)) Let $V = \{v_1, v_2, \ldots, v_m\}$ be a set of vertices in any graph, G, and $E = \{e_1, e_2, \ldots, e_n\}$ be a set of edges in G. Then, *MES* values for each element in V and E can have user-specified values or other ones discussed in [4]. $MES(v_k)$ and $MES(e_l)$ mean *MES* values for v_k and e_l respectively. If a support of any element x, $SUP(x)$ is lower than $MES(x)$, x becomes useless one.

Graph patterns are composed of a number of the elements, and each element has its own *MES* value. General graph mining methods determine with a given minimum

support threshold whether any graph pattern is frequent or not. However, in this paper, we use a different criterion to solve the *rare item problem*.

Definition 2. (*MGS*(*Minimum sub-Graph Support*)) Given a sub-graph, G, its $V = \{v_1, v_2, \ldots, v_i\}$, and $E = \{e_1, e_2, \ldots, e_j\}$, then, $MGS(G) = minimum(MES(v_1), \ldots, MES(v_i), MES(e_1), \ldots, MES(e_j))$. *MGS* is used to check whether G is valid or not. Thus, if $SUP(G)$ is smaller than $MGS(G)$, G becomes an invalid sub-graph pattern.

The reason why *MGS* is computed in this way is to consider the rarest element among the graph. Therefore, a certain sub-graph satisfying the *MGS* condition becomes a valid pattern with useful rarity. However, if we directly delete elements or sub-graphs which do not satisfy the *MES* or *MGS* conditions, fatal pattern losses can be caused since pruning by them does not satisfy the anti-monotone property. That is, given a sub-graph with invalid elements, although it is currently infrequent, certain supper patterns of it may be frequent as its graph expansion works are performed. In all of the pattern mining areas, maintaining this property is one of the most important factors since we can prune invalid elements or sub-graph patterns in advance only when that property is effective. For this reason, we need a way to maximize pruning efficiency maintaining that property.

Definition 3. (*LMS*(*Least Minimum Support for the graph mining*)) *LMS* is a pruning factor without causing any pattern loss. Given a set of *MES*, S_{MES}, it is sorted in *MES* descending order, denoted as $S_{MES} = \{MES(x_i), MES(x_j), \ldots, MES(x_l)\}(\forall x \mid x \in V \text{ or } E, MES(x_i) \geq MES(x_j) \ldots \geq MES(x_l))$. After that, Starting from x_l, we compare $SUP(x_l)$ with $MES(x_l)$. If there exists any element, x_k, satisfying $SUP(x_k) \geq MES(x_k)$, $MES(x_k)$ becomes *LMS*. Pruning works by *LMS* maintains the anti-monotone property.

Certain elements having lower supports than *LMS* do not make valid sub-graphs in any case since all possible super patterns expanded from them have also lower values than it. Likewise, if a certain graph has a smaller support than *LMS*, its all of the possible supper sub-graph patterns also have lower values than *LMS*. Accordingly, they can be pruned in advance without causing any problem. That is, the anti-monotone property is satisfied. In addition, by pruning those useless ones, we can reduce runtime and memory resources needed for the mining operations effectively.

3.2 FGM-MMS Algorithm

Fig. 1 presents an overall mining procedure of our FGM-MMS. In the procedure: *FGM-MMS*, the algorithm first calculates *LMS* according to the definition 3 and then finds valid V and E(lines 1~2). After that, it generates valid sub-graphs considering the multiple minimum support constraints, performing expansion operations for graphs(lines 3~6). When the sub-procedure, *Subgraph-expansion* is called, the algorithm conducts the graph expansion works with different ways depending on current graph forms(lines 2~3). If a support of the expanded graph, G', $SUP(G')$ is lower than *LMS*, it is permanently removed. Otherwise, the support is compared with $MGS(G')$ again, where G' is inserted in FG if its support is not smaller than $MGS(G')$(line 5). Thereafter, the algorithm performs mining operations expanding G' recursively(lines 6~7). After all of the expansions terminate, we can gain the complete FG considering the multiple minimum support constraints.

Input GDB : a given graph database S_{MES} : a set of Minimum element support values
Output FG : a set of frequent sub-graph patterns

Procedure: FGM-MMS(GDB, S_{MES})
01. Calculate LMS from S_{MES} //according to the definition 3
02. V, E ← vertices and edges such that their supports ≥ LMS;
03. For each vertex, v_i in V, do {
04. $G \leftarrow v_i$; E' ← a set of edges that can be attached to v_i in E;
05. $FG \leftarrow FG \cup$ *Subgraph-expansion*(G, E');
06. } //end for

Sub-procedure: Subgraph-expansion(G, E)
01. For each edge, e_i in E, do {
02. If (G = *path* or *free tree*) : G' ← adding to G both e_i and the vertex contained in e_i;
03. Else : G' ← inserting to G only e_i such that it is a cyclic edge; //the case of a cyclic graph
04. If ($SUP(G') \geq LMS$)
05. If ($SUP(G') \geq MGS(G')$: $FG \leftarrow FG \cup G'$;
06. E' ← edges that can be attached to G';
07. $FG \leftarrow FG \cup$ *Subgraph-expansion*(G', E');
08. } //end for

Fig. 1. FGM-MMS algorithm

4 Performance Analysis

We compare our FGM-MMS with Gaston[5, 6] for objective performance evaluation since the Gaston, as a state-of-the-art algorithm, has more outstanding performance than the others. They were written in C++ and ran with 3.33GHz CPU, 3GB RAM, and WINDOWS 7 OS. We use a real graph dataset, named DTP[6], to show how effectively our FGM-MMS is performed on the real-world graph data.

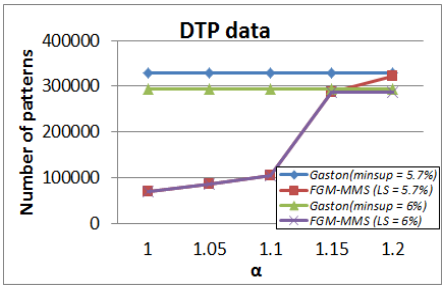


Fig. 2. The number of frequent sub-graph patterns

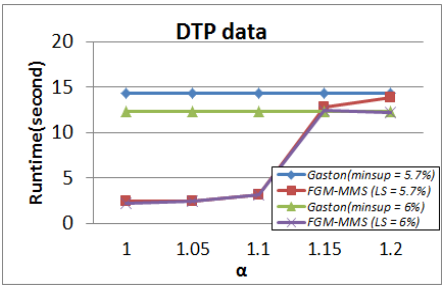


Fig. 3. Runtime result

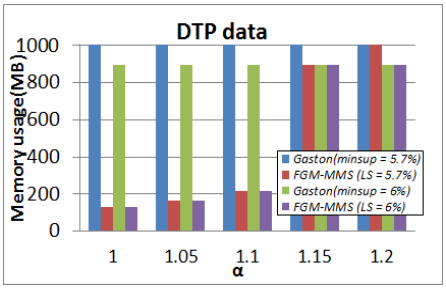


Fig. 4. Memory usage

To assign *MES* for each element, we applied the methodology discussed in [4]. Its formula is $MES(e_k) = maximum(\beta * SUP(e_k), LS)$, where $SUP(e_k)$ means a support of any element, e_k, and *LS* is the user-specified lowest *MES* and has the same value as the minimum support threshold used in Gaston. $\beta(0 \leq \beta \leq 1)$ represents how closely *MES* for a certain element is related to its support. As the value becomes closer to 1, relevance becomes higher. It is denoted as $\beta = 1/\alpha$, and α is variable.

In Fig. 2 to 4, we can observe that FGM-MMS shows outstanding performance more than that of Gaston in all cases when α is relatively low. The reason is as follows. When α is high, *MES* values are more likely to be set as *LS*, so in this case, our algorithm operates similarly to Gaston. Meanwhile, at lower values of α, FGM-MMS can better reflect support values for each element to their own *MES*, which means that the effect of the multiple minimum supports becomes larger. Thus, the proposed algorithm can extract valid sub-graph patterns with smaller runtime and memory resources as shown in the figures.

5 Conclusion

In this paper, we proposed a frequent graph mining algorithm based on multiple minimum support constraints, called FGM-MMS. In contrast to the previous algorithms using a single minimum support threshold, our algorithm could solve the challenging issue, *rare item problem* on the graph mining area by using the multiple minimum support constraints, thereby mining more meaningful frequent sub-graph patterns. Furthermore, we demonstrated that the proposed algorithm shows more outstanding performance compared to the previous state-of-the-art algorithm through comprehensive performance experiments provided in this paper.

References

1. Hu, Y.H., Chen, Y.L.: Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. Decision Support Systems 42(1), 1–24 (2006)
2. Kiran, R.U., Reddy, P.K.: Novel techniques to reduce search space in multiple minimum supports-based frequent pattern mining algorithms. In: 14th International Conference on Extending Database Technology, pp. 11–20 (2011)
3. Lee, G., Yun, U.: Mining weighted frequent sub-graphs with weight and support affinities. In: Proceedings of the 6th Multi-Disciplinary International Workshop on Artificial Intelligence, pp. 227–238 (2012)
4. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 337–341 (1999)
5. Nijssen, S., Kok, J.N.: The Gaston Tool for Frequent Subgraph Mining. Electronic Notes in Theoretical Computer Science 127(1), 77–87 (2005)
6. Nijssen, S., Kok, J.N.: A quickstart in frequent structure mining can make a difference. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 647–652 (2004)

Design of Automatic Paper Identification System with QR Code for Digital Forensics

Ha-Kyung Jennifer Lee[1,*], Young-Mi Yun[2], Kee-Hyung Yoon[2], and Dong-Sub Cho[1]

[1] Dept. of Computer Science and Engineering, Ewha Womans University, Seoul, Korea
lhakyungl@gmail.com, dscho@ewha.ac.kr
[2] Document Examination Section, Supreme Prosecutors' Office, Seoul, Korea
{nymph,keehyung}@spo.go.kr

Abstract. As the printing technology is developing, the use of digital documents which is printed with a variety of printing papers is increasing. And the increase in criminal forgery of digital documents has been the cause of many social problems. As a result, the development of digital forensic techniques to collect and analyze the evidence of the crime of forgery of digital documents is increasingly important. In this paper, we propose an automatic paper identification technique with QR code that is formed based on the extracted features from each paper's microscope image.

Keywords: Printing Paper, Paper Id, Identification System, QR code, Digital Forensics, Digital Document.

1 Introduction

Nowadays using a computer is common in most age groups, and various documents are written on a computer and printed by various printers. Unlike the handwritten documents that have personalized handwriting features making those documents easily distinguished, the digital documents printed by printer are difficult to distinguish the authors who have made the documents.

To identify the print media such as printing paper that is used to print out the document to determine whether the printed digital document is forged will be play a very important role in the field of digital document forensics.

In this paper, high-resolution optical microscope digital images of printing papers were used to extract the feature values of each printing papers to identify the printing paper. The digital images were obtained by magnifying the specific location of each printing papers 500 times (monitor-magnification) and they were transformed by using image processing techniques such as grayscale and color quantization, and then we extracted the value of the printing paper's unique features. The value became the input string of QR code generator and we could get the QR code to identify the printing paper.

Through this study, it was able to implement an automatic identification system to distinguish printing paper using QR code.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_5,

This paper is organized as follows. In Section 2, we explain the design of the proposed automatic paper identification system and show each stage in detail. In Section 3, the output of the proposed system is shown, and finally in Section 4, we come to a conclusion.

2 Automatic Paper Identification System with QR Code

2.1 Proposed System for Paper Identification

In the field of digital forensics, the research on how to determine whether the digital document is forged has been constantly evolving, and the research using the unique features of the fibers of the printing paper has been significantly addressed in the field of digital forensics. Because of randomly generated paper fiber distribution features in the manufacturing process of paper, the features are almost impossible to replicate and this may be a unique feature that can determine whether the printed digital documents are forged or not. To prevent the forgery of digital documents, the extraction technology of the printing paper's unique characteristics of the distribution of paper fibers can be very useful in the field of digital forensics.

In this study, we describe a technique to distinguish between each of the printing paper through a unique feature values extracted from the digital image which is magnified at specific area of the paper by the high-resolution digital optical microscope. 100 different printing papers were used as input for the automatic Paper Identification system, and high-resolution images using an optical microscope is magnified 500 times (monitor magnification).

The major processes of the system are displayed in Fig. 1 as follows:

- Stage 1: Image Processing
- Stage 2: Extracting Feature Value of Paper
- Stage 3: Generating QR code

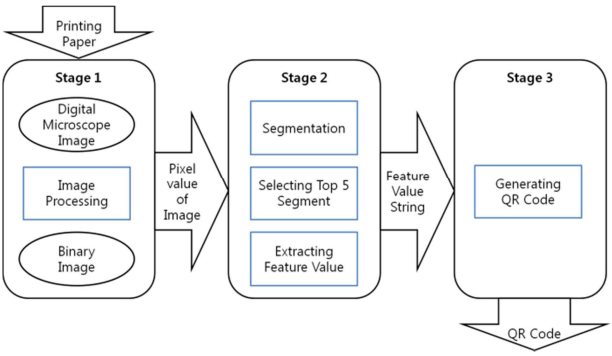


Fig. 1. Architecture of Automatic Paper Identification System

2.2 Automatic Paper Identification System Modules

To obtain the feature value of each printing paper, the paper is photographed at a monitor magnification of 500, and the image captured is color photograph with 800×600 pixels in BMP format.

(1) Input digital image goes through the process of image preprocessing by gray-scale and color quantization techniques (Fig. 2).

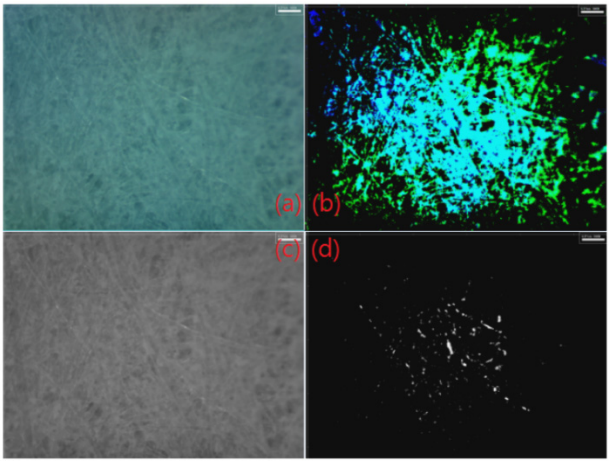


Fig. 2. (a) original image input, (b) color quantization image, (c) gray-scaled image, (d) gray-scaled color quantization image

Image processing is an action to transform the two-dimensional picture and the gray scale technique and color quantization technique are widely used method of image processing techniques.

The gray scale technique is to express the image of white, black, and 256 different colors including the multi-step gray. It uses the properties of color that color can be represented by the gray scale if red, green, blue values of the color are equal. With this technique, we can express more sophisticated image which consists of only brightness, and the equation is as follows:

$$\text{Gray} = (0.2989 * \text{Red}) + (0.5870 * \text{Green}) + (0.1140 * \text{Blue}) \quad (1)$$

Color quantization technique is a technique that creates entirely new image by reducing the number of colors used in the original image using distinct colors to look similar to the original image. The image processing procedure is as follows:

Procedure of Color Quantization [2]

1) Sampling the original image for color statistics
2) Choosing a color-map based on the color statistics (frequency distribution)
3) Mapping original colors to their nearest neighbors in the color-map
4) Quantizing and redrawing the original image

(2) Divide area according to the distribution of colors in the image (Fig. 3).

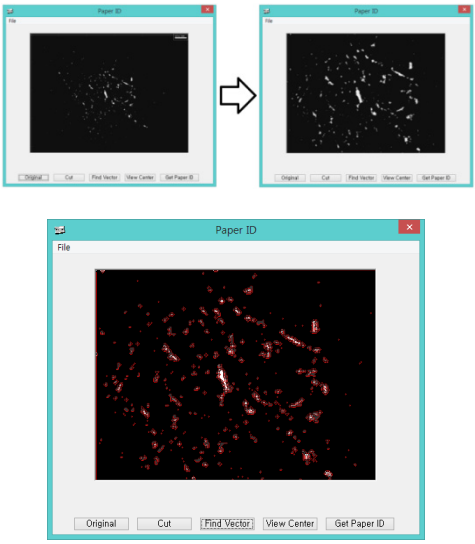


Fig. 3. Segmenting the image after enlarge the center area of the image

(3) According to the order of size of the segmented area, calculate the center of gravity of top five largest areas.

$$Area = A = \frac{1}{2}\sum_{i=0}^{N-1}(x_i y_{i+1} - x_{i+1} y_i)$$

$$Centroid_x = \frac{1}{6A}\sum_{i=0}^{N-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

$$Centroid_y = \frac{1}{6A}\sum_{i=0}^{N-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

Fig. 4. The equations of polygon's area and the center of gravity

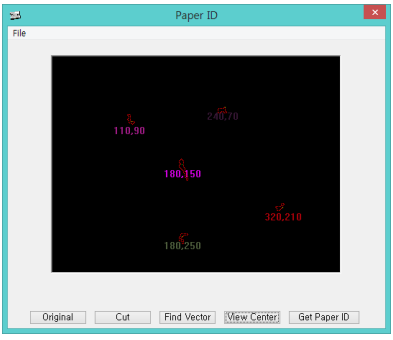


Fig. 5. Calculated Center Coordinates of Top 5 Areas

(4) Calculate the ratio of the width and height of the smallest rectangle consisting of the five selected coordinates. And obtain the relative position of the five selected coordinates.

Feature Value 1: width / height of the rectangle (2)

Feature Value 2: relative position of the five selected coordinates (3)

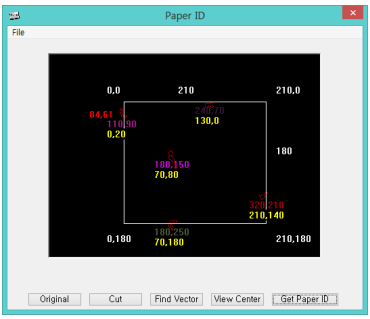


Fig. 6. Calculating feature values

(5) Get QR Code image with the feature value string (Fig. 7).

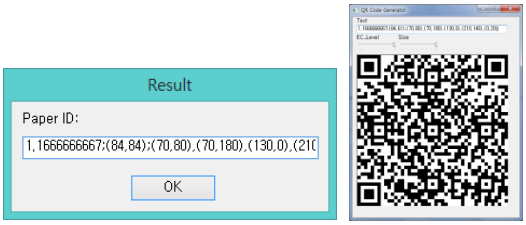


Fig. 7. The Result of QR Code Image of Paper

3 Results

Through the proposed system, every printing paper can have unique QR code image. The results are shown in Table 1.

Table 1. Results

Paper no.	Paper_ID (feature value)	QR code
1	1.1666666667;(84,61); (70,80),(70,180),(130,0),(210,140),(0,20)	

Table 1. (*continued*)

2	0.7222222222;(29,58); (130,0),(70,10),(10,180),(0,40),(70,30)	
3	1.6470588235;(154,56); (90,10),(220,100),(0,0),(280,10),(220,170)	
4	1.4444444444;(62,41); (130,90),(0,40),(10,0),(130,40),(80,60)	
5	0.8750000000;(58,74); (140,100),(0,100),(130,150),(0,0),(90,160)	

4 Conclusions

Through this study, it is now possible to determine the source of the digital document using the Paper ID QR code image. If the encrypted QR code image is marked on each paper, it will be possible to determine more quickly whether the printed digital document is forged. Also if we standardize the unique features of printing paper, we can expect that this study is useful in the field of digital forensics in the future.

References

1. http://en.wikipedia.org/wiki/Grayscale
2. Heckbert, P.S.: Color Image Quantization for Frame Buffer Display. In: ACM SIGGRAPH 1982 Proceedings of the 9th Annual Conference on Computer Graphics and Interactive Techniques, pp. 297–307 (1982)
3. Yamakoshi, M., Tanaka, J., Furuie, M., Hirabayashi, M., Matsumoto, T.: Individuality evaluation for paper based artifact-metrics using transmitted light image. In: Proc. SPIE 6819, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, 68190H (2008)

Processing Continuous Range Queries with Non-spatial Selections

HaRim Jung[1], Seongkyu Kim[1], Joon-Min Gil[2], and Ung-Mo Kim[1]

[1] School of Information and Communication Engineering, Sungkyunkwan University, Suwon, Korea
[2] School of IT Engineering, Catholic University of Daegu, Gyeongbuk, Korea

Abstract. In this paper, we explore the problem of scalable evaluation of Continuous Range Queries (CRQs) with non-spatial selections, each of which continually retrieves the moving objects that (i) are currently within a specified spatial query region and (ii) satisfy specified non-spatial selections. We propose a new query indexing structure, called the Bit-vector Query Region tree (BQR-tree), which enables the server to cooperate with moving objects for evaluation of CRQs with non-spatial selections. Through simulations, we verify the efficiency of the BQR-tree.

Keywords: Continuous range queries, moving objects, index structures.

1 Introduction

Many useful Location Based Services (LBSs) usually rely on the functionality of evaluating *Continuous Range Queries* (*CRQs*), each of which continually retrieves the moving objects that are currently located within a spatial query region of interest a client specifies. Consider the following scenario of a location based advertising service as an example, where a restaurant (i.e., client) plans to send e-discounts to the nearby potential customers (i.e., moving objects). Then, the service provider (i.e., server) must be able to keep track of the locations of these customers and report them to the restaurant, whenever needed. In many real-life LBSs, however, clients of diverse interests often additionally constrain the target moving objects by specifying their non-spatial selection criteria. Continuing with the above example, suppose the restaurant focuses on attracting only a specific class of customers (e.g., $30 \leq Age \leq 40$, \$40, 000 $\leq$ *Annual income* $\leq$ \$80, 000, and *Dietary preference* = Vegetarian). In this case, the service provider should report only the nearby customers whose profiles match the above criteria to the restaurant.

The majority of existing methods for CRQ evaluation assumed that moving objects periodically send location-updates to the server via wireless connections and the server keeps the results of the registered queries up-to-date [1]. However, in case the server involves a large number of moving objects, the overall system performance will deteriorate drastically due to the overwhelming server workload and severe communication bottleneck. To help the moving objects reduce the frequency of

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_6,

sending location-updates, the *safe region* method was introduced in [2]. The safe region, assigned to each moving object o, is the area that (i) contains o and (ii) guarantees the current results of all the queries to remain valid as long as o does not exit it. Therefore, o need not send its location-update to the server until it does not exit its safe region (See o_1 in Fig. 1).

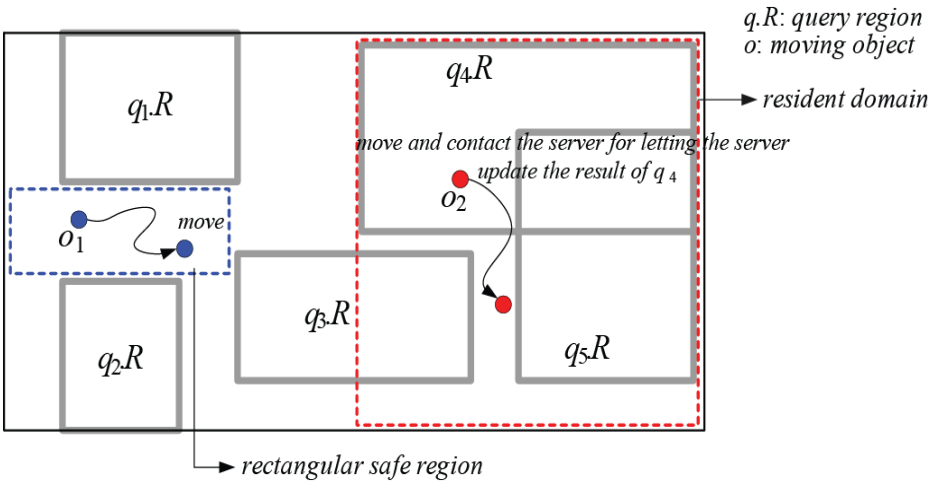


Fig. 1. An example of the safe region and resident domain

Monitoring Query Management (*MQM*), which aims to reduce the server workload and communication cost by leveraging the available capabilities of moving objects, was introduced in [3]. In MQM, the server assigns each moving object o (i) a subdomain of the entire workspace, called the *resident domain*, which contains o and (ii) several query regions that overlap with the resident domain. The size of o's resident domain is determined by o's capability, which is measured by the maximum number of query regions o can load and process at a time. Only when each moving object o exits its resident domain or crosses any of its assigned query regions, it contacts the server to receive a new resident domain (together with new query regions) or let the server update the query results (See o_2 in Fig. 1). In [4], the *Space Partitioning Query Index* (*SPQI*) was introduced to improve the performance of MQM.

All the methods reviewed above cannot adequately deal with CRQs with selections on non-spatial attributes because they only consider spatial attributes. In this paper, we use the resident domain concept and propose a novel query indexing structure, referred to as the *Bit-vector Query Region tree* (*BQR-tree*), which enables the server to cooperate with moving objects for efficient evaluation of CRQs with non-spatial selections. The BQR-tree indexes queries based on query regions, and it augments each node with non-spatial information in the form of bit-vector. Through simulations, we verify the superiority of the BQR-tree against existing methods.

2 The Proposed Method

2.1 Problem Formulation and Motivation

Problem Formulation. Suppose each moving object o is associated with a set of n non-spatial attributes $A = \{a_1, a_2, \cdots, a_n\}$. Each non-spatial attribute a_i in A ($1 \le i \le n$) is either numeric (e.g., *Age* and *Annual income*) or categorical (e.g., *Dietary preference*). We denote the non-spatial attribute values of o by $o.A = \{o.a_1, o.a_2, \cdots, o.a_n\}$. A query q is represented as $(q.R, q.V)$, where $q.R$ denotes a spatial query region and $q.V = \{q.v_1, q.v_2, \cdots, q.v_m\}$ denotes a set of non-spatial intervals or non-spatial values specified on a subset of non-spatial attributes $\acute{A}$ $(\subseteq A) = \{\acute{a}_1, \acute{a}_2, \cdots, \acute{a}_m\}$. We assume in this paper that $q.v_m$ is an interval in case $\acute{a}_i$ is a numerical attribute. For a query $q = (q.R, q.V)$, the server should continually retrieve all moving objects o that are currently located within $q.R$ and satisfy the following: $\forall o.\acute{a}_i \in o.\acute{A}$, $o.\acute{a}_i$ lies in $q.v_i$ (or $o.\acute{a}_i = q.v_i$ if $\acute{a}_i$ is a categorical attribute).

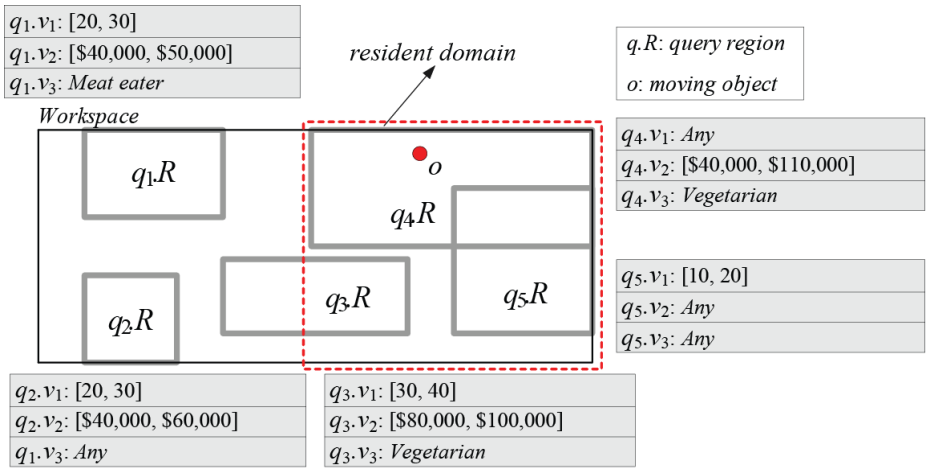


Fig. 2. Spatial query regions with non-spatial selections

Motivation. Assuming each moving object o has some available (memory and computational) capability $o.Cap$, the goal of our work is to make the server and moving objects share the evaluation process of queries with non-spatial selections. To this end, we use the resident domain concept. Let us assume that the moving object o with $o.Cap$[1] = 3 in Fig. 2 is associated with three non-spatial attributes $A = \{a_1$: *Age*, a_2: *Annual income*, a_3: *Dietary preference*$\}$ and $o.A = \{o.a_1 = 35, o.a_2 = \$93,000, o.a_3 =$ *Vegetarian*$\}$. Suppose the queries $q_1 \sim q_5$ involve non-spatial selections $q_1.V \sim q_5.V$ specified on a subset of A as shown in Fig. 2 in addition to the query regions $q_1.R \sim q_5.R$. In MQM and SPQI, the server assigns o the resident domain depicted in the figure together with the query regions $q_3.R$, $q_4.R$, and $q_5.R$. However, o can be assigned

[1] In this paper, we assume that $o.Cap$ is measured by the maximum number of query regions o can load and process at a time.

a much larger resident domain than its current resident domain (e.g., the whole workspace) because o does not satisfy the non-spatial selection criteria the queries q_1, q_2, and q_5 specify. Therefore, o need not check its movement against the query regions $q_1.R$, $q_2.R$, and $q_5.R$ because its movement does not affect the results of q_1, q_2, and q_5. This clearly helps o reduce the frequency of contacting the server for receiving new resident domains or updating the query results. We say a query region $q.R$ is *non-spatially matched* to a moving object o if o satisfies the non-spatial selection criteria the corresponding query q specifies.

2.2 The Bit-vector Query Region Tree (BQR-Tree)

The Bit-vector Query Region tree (BQR-tree) indexes queries based on spatial query regions, and it augments each node with non-spatial information in the form of bit-vector. We classify the overlap relationship between a spatial query region $q.R$ and a (sub) domain N into four cases as shown in Fig. 3: *covers* (Fig. 3a), *is covered by* (Fig. 3b), *partially intersects* (Fig. 3c), and *equals to* (Fig. 3d).

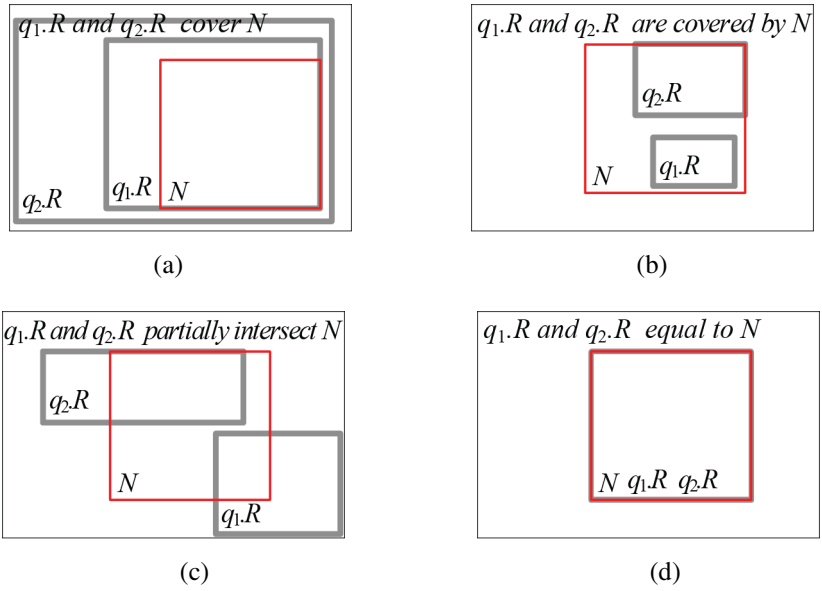


Fig. 3. Classification of the overlap relationship

We represent non-spatial attribute values of moving objects and non-spatial intervals (or values) of queries as *object bit-vectors* and *query bit-vectors*, respectively. For a numerical attribute a, its mapping function f divides its domain into $|IV|$ disjoint intervals iv_1, iv_2, ⋯, $iv_{|IV|}$ of equal length. Then, given a moving object o, f maps $o.a$ into a bit-string $(b_1, b_2, \cdots, b_{|IV|})$ such that $b_i = 1$ if $o.a$ lies in iv_i, otherwise $b_i = 0$. Similarly, it maps the interval $q.v$ a query q specifies on a into $|IV|$ bit-string $(b_1, b_2, \cdots, b_{|IV|})$ such that $b_i = 1$ if $q.v$ overlaps iv_i, otherwise $b_i = 0$. On the other hand,

for a categorical attribute $\acute{a}$ with $|C|$ categories c_1, c_2, ⋯, $c_{|C|}$, a mapping function $\acute{f}$ maps $o.\acute{a}$ (or the specified value $q.\acute{v}$ on $\acute{a}$) into $\acute{v}$ a bit-string (b_1, b_2, ⋯, $b_{|C|}$) such that $b_j = 1$ if $o.\acute{a} = c_j$ (or $q.\acute{v} = c_j$), otherwise $b_j = 0$.

Definition 1. ***Object bit-vector***: *Suppose that there is a mapping function f_i ($1 \le i \le n$) for each non-spatial attribute $a_i \in A$ ($1 \le i \le n$). Then, an object bit-vector generated for o.A is $f_1(o.a_1) + f_2(o.a_2) + \cdots + f_n(o.a_n)$, where + denotes the bit-string concatenation operator.*

Definition 2. ***Query bit-vector***: *A query bit-vector generated for q.V is $f_1(q.v_1) + f_2(q.v_2) + \cdots + f_n(q.v_n)$. Note that in case q.V does not contain the specified interval (or value) $q.v_i$ on a_i ($1 \le i \le n$), the bit-string for $f_i(q.v_i)$ becomes ** ⋯ * with its length being equal to $f_i(o.a_i)$, where the symbol * denotes a don't care condition.*

In the following, we show an example of generating an object bit-vector and query bit-vectors using the non-spatial attribute values *o.A* of the moving object *o* in Fig. 2 and the sets of non-spatial selections $q_4.V$ and $q_5.V$ the queries q_4 and q_5 specify on a subset of non-spatial attributes *A* in Fig. 2. Suppose that there are three mapping functions as follows:

$$f_1(x) = \begin{cases} 1000 \text{ if } x \text{ lies in (or overlaps) } [0, 20); \\ 0100 \text{ if } x \text{ lies in } [20, 40); \\ 0010 \text{ if } x \text{ lies in } [40, 60); \\ 0001 \text{ otherwise.} \end{cases}$$

$$f_2(x) = \begin{cases} 1000 \text{ if } x \text{ lies in (or overlaps) } [0, 40000); \\ 0100 \text{ if } x \text{ lies in } [40000, 80000); \\ 0010 \text{ if } x \text{ lies in } [80000, 120000); \\ 0001 \text{ otherwise.} \end{cases} \qquad f_3(x) = \begin{cases} 10 \text{ if } x \text{ is Meat eater;} \\ 01 \text{ if } x \text{ is Vegetarian.} \end{cases}$$

Then, the object bit-vector generated for *o.A* is 0100001001 (= 0100 + 0010 + 01). On the other hand, the query bit-vectors generated for $q_4.V$ and $q_5.V$ are ****001001 (=**** + 0010 + 01) and 1000 ****** (= 1000 + **** + **), respectively. Now, we present the details of the BQR-tree. The BQR-tree is a binary tree index of queries, which is built by recursive split of the entire workspace. Given a set of query regions on the workspace that corresponds to the root, if the number of these query regions is greater than the split threshold *t*, it is split into two subdomains, each of which corresponds to a child node of the root. This process recursively continues until every subdomain has no more than *t* query regions that are covered by or partially intersect the subdomain, and it corresponds to a leaf node. The threshold value *t* is determined by the moving object with the minimum capability among all the moving objects registered at the server.

A leaf node of the BQR-tree stores at most *t* entries of the form (*qid*, *q.bv*), where *qid* refers to a query *q* in the query table and *q.bv* is the query bit-vector of *q*. A non-leaf node stores two entries of the form (*ptr*, *N*, *N.bv*), where *ptr* is a pointer to a child node (i.e., leaf or non-leaf node), *N* is a subdomain of the child node pointed to by *ptr*, and *N.bv* is the node bit-vector of a node *N*.[2] Each node of the BQR-tree stores a

[2] Hereafter, without ambiguity, we use the symbol '*N*' to denote both a tree node and its corresponding (sub) domain.

variable *Count* and is associated with a covering list *CL*. The BQR-tree satisfies the following properties:

1. An entry (*qid*, *q.bv*) for a query *q* is stored in a leaf node *N* only if *q.R* is covered by or partially intersects *N*.
2. An entry (*qid*, *q.bv*) for a query *q* can be redundantly stored in several leaf nodes if *q.R* partially intersects these leaf nodes.
3. For each entry (*ptr*, $\acute{N}$, $\acute{N}.bv$) stored in a non-leaf node *N*, $\acute{N}$ represents one of the equal halves of *N*'s domain, whereas $\acute{N}.bv$ is a node bit-vector of $\acute{N}$, which is formed by bitwise OR-ing every query bit-vector *q.bv* of each query *q* whose query region *q.R* is covered by or partially intersects $\acute{N}$.
4. For each (leaf or non-leaf) node *N*, *N.Count* records the total number of query regions that are covered by or partially intersect *N*.
5. For each (leaf or non-leaf) node *N*, its associated covering list *N.CL* keeps every entry (*qid*, *q.bv*) for each query *q* whose query region *q.R* covers or equals to *N*.

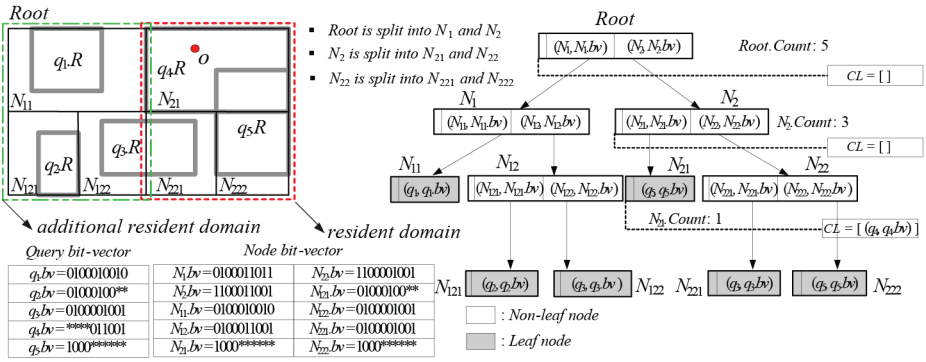


Fig. 4. An example of the BQR-tree

Assuming $t = 1$, Fig. 4 shows the BQR-tree for the queries $q_1 = (q_1.R, q_1.V)$, $q_2 = (q_2.R, q_2.V)$, ⋯, $q_5 = (q_5.R, q_5.V)$ in Fig. 2. For example, the query regions $q_4.R$ and $q_5.R$ are covered by the node N_2, and the query region $q_3.R$ partially intersects N_2. Therefore, the node bit-vector $N_2.bv$ of N_2 is 100011001 (= $q_3.V \vee q4.bv \vee q5.bv$, where $\vee$ denotes bit-wise OR-ing). When a new moving object is registered at the server, the search algorithm for its resident domain is invoked. Fig. 5 is the pseudo code of the search algorithm on the BQR-tree for assigning each moving object its resident domain *N* together with non-spatially matched query regions that are covered by or partially intersect *N*. Due to the space limit, we omit the details.

After the search algorithm terminates, the server searches all the queries (in the query table) referred to by the retrieved entries and assigns the moving object *o* (i) its resident domain *N*, (ii) the pairs of non-spatially matched query region and the entry for the corresponding query, and (iii) the additional resident domain (if possible). For example, the moving object *o* with *o.Cap* = 3 in Fig. 4 is assigned (i) the node N_2 as its resident domain, (ii) the pairs of non-spatially matched query region and the entry for the corresponding query ((q_3, $q_3.bv$), $q_3.R$), ((q_4, $q_4.bv$), $q_4.R$), and (iii) the

additional resident domain N_1 (because N_1 is the topmost element in *Queue*, which spatially touches N_2). The cooperation protocol between moving objects and the server is quite similar to that in [4]. Therefore, we omit it.

Algorithm. SEARCH(N, o)
Input N: A QR-tree node initially set to the root, o: A moving object
Output R: o's resident domain, E: o's additional resident domain
QE: A set of distinct qualifying entries
1: Map $o.A$ to $o.bv$;
2: Initialize a queue *Queue*;
3: **for** each entry (ptr, $\acute{N}$) stored in N **do**
4: **if** $\acute{N}$ contains o's current location **then** // $\wedge$ denotes bit-wise AND-ing
5: **if** $\acute{N}.bv \wedge o.bv \neq o.bv$ **then**
6: Set R to $\acute{N}$; // Because $\acute{N}$ stores no query regions that are non-spatially matched to o
7: **else if** $\acute{N}.bv \wedge o.bv = o.bv$ **then**
8: **if** $\acute{N}.Count \leq o.Cap$ **then**
9: Set R to $\acute{N}$;
10: $QE \leftarrow QE \cup$ FINDQUERYREGIONS;
11: **else**
12: SEARCH($\acute{N}$, o); // Recursion
13: **else** // $\acute{N}$ does not contain o's current location
14: Enqueue $\acute{N}$ if $\acute{N}$.bv $\wedge$ $o.bv \neq$ o.bv;
15:**if** $|QE| < o.Cap$ **then**
16: **repeat**
17: Dequeue the next element E from Queue; // Finds o's additional resident domain
18: **until** E spatially touches R or *Queue* is empty

Note: FINDQUERYREGIONS is a simple depth-first search algorithm that takes $\acute{N}$ as an input and retrieves all the distinct query identifiers stored in each $\acute{N}$'s descendent leaf node $\ddot{N}$ and its associated covering list $\ddot{N}.CL$.

Fig. 5. Pseudo code of the search algorithm

3 Performance Evaluation

This section compares the performance of the BQR-tree method (denoted by BQRT) with the safe region method (denoted by SR), MQM, and SPQI with regard to the server workload and communication cost. The server workload was measured in terms of the amount of CPU-time the server takes for CRQ evaluation. On the other hand, the communication cost was measured by the sum of (i) the number of location-updates sent from moving objects to the server and (ii) the number of messages sent from the server to the moving objects. We set query regions to be uniformly placed on the workspace. The movements of the moving objects we generated follow the *random waypoint model*. The computational capability of each moving object was randomly selected from the range between 25 and 100 query regions, and thus the threshold value t of the BQR-tree, BP-tree (used in MQM) and SPQI was set to 25. Each moving object is associated with five non-spatial attributes a_1, a_2, ⋯, a_5, where a_1, a_2 and a_3 are numerical while a_4 and a_5 are categorical. The distribution of each non-spatial attribute value $o.a_1$, $o.a_2$, ⋯, $o.a_5$ of each moving object follows the *Zipf* distribution with skew coefficient $\alpha = 0.8$. Each non-spatial interval or value specified

on a subset of a_1, a_2, ⋯, a_5 by each query q follows the same distribution. We ran 1,000 simulation time steps and measured the average of the CPU-time (in millisecond) and the average of the total number of messages by varying the number of moving objects from 20,000 to 200,000. As shown in Fig. 6, the overhead of all the methods increases in terms of the amount of CPU-time and the number of messages as the number of moving objects increases. However, BQRT outperforms SR, MQM, and SPQI due to the fact that only BQRT has the ability to fully utilize the capabilities of moving objects by ignoring the query regions that are not non-spatially matched to the moving objects.

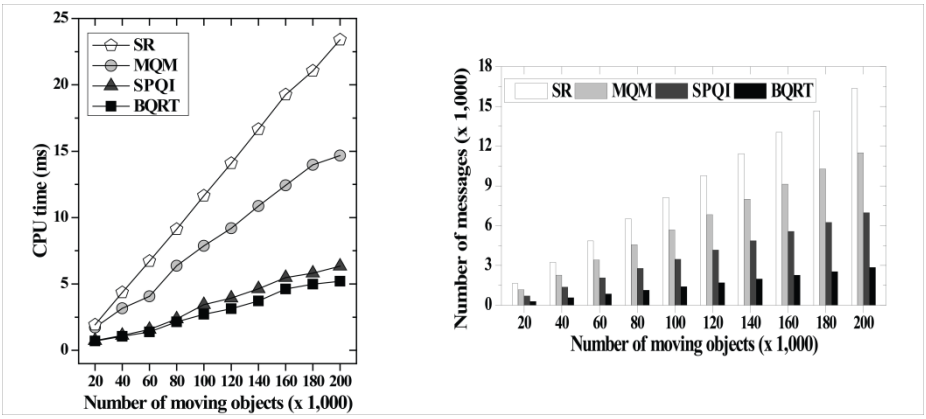


Fig. 6. CPU-time and # of messages vs. # of moving objects

4 Conclusions

In this paper, we presented the Bit-vector Query Region tree (BQR-tree) for evaluating continuous range queries with non-spatial selections. The BQR-tree, which (i) stores the additional bit-vector information to describe the non-spatial information, and (ii) enables the server to cooperate with moving objects for evaluation of queries, greatly improves the overall system performance when the queries involve non-spatial selections.

References

1. Ilarri, S., Mena, E., Illarramendi, A.: Location-Dependent Query Processing:Where We Are and Where We Are Heading. ACM Computing Surveys 42(3), 1–73 (2010)
2. Hu, H., Xu, J., Lee, D.: A Generic Framework for Monitoring Continuous Spatial Queries over Moving Objects. In: Proc. of ACM SIGMOD, pp. 479–490 (2005)
3. Cai, Y., Hua, K.A., Cao, G., Xu, T.: Real-Time Processing of Range-Monitoring Queries in Heterogeneous Mobile Databases. IEEE Trans. on Mobile Computing 5(7), 931–942 (2006)
4. Jung, H., Kim, Y.S., Chung, Y.D.: SPQI: An Efficient Index for Continuous Range Queries in Mobile Environments. Journal of Information Science and Engineering 29(3), 557–568 (2013)

DSPI: An Efficient Index for Processing Range Queries on Wireless Broadcast Stream*

Kwanho In, Seongkyu Kim, and Ung-Mo Kim

School of Information and Communication Engineering,
Sungkyunkwan University, Suwon, Korea

Abstract. This paper addresses the problem of processing range queries on wireless broadcast streams. In order to support range queries efficiently, we propose a novel index called Distributed Space-Partitioning Index (DSPI). DSPI consists of hierarchical grids that provide mobile clients with the global view as well as the local view of the broadcast data. The algorithm for processing range queries based on DSPI is also proposed. Simulation experiments demonstrate DSPI is superior to the existing index schemes.

Keywords: Continuous range queries, moving objects, index structures.

1 Introduction

With the advances in wireless technologies and wide spread mobile devices, the trend toward pervasive computing has gained momentum. Since location is one of the most important properties in a pervasive computing environment, various location dependant information services (LDISs) have emerged as one of the promising applications [3, 4]. In order to support LDISs, efficient processing of location dependant queries (LDQs), which retrieve information based on the current locations of mobile clients (MCs), is critical. Wireless data broadcast is considered to be an effective way for disseminating location dependant data (LDD) and supporting LDQs since it leverages computational capability of MCs, and thus accommodates a huge number of MCs simultaneously. In this paper, we address the problem of processing range queries (one of the essential classes of LDQs) in wireless broadcast environments.

In wireless data broadcast, data (e.g., LDD) are periodically broadcasted. MCs then may retrieve the data and evaluate their queries (e.g., LDQs) on the broadcast stream. In order to retrieve necessary data for query processing, MCs have to continuously listen to the broadcast channel until the desired data arrive. This leads to vast *energy consumption* on MCs. To alleviate such energy consumption, the concept of air index is introduced. With the index information, MCs can read only the required data selectively. However, the *waiting time* of the MCs is increased because the insertion of index information extends the *broadcast cycle*. Two performance measures are

* This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.

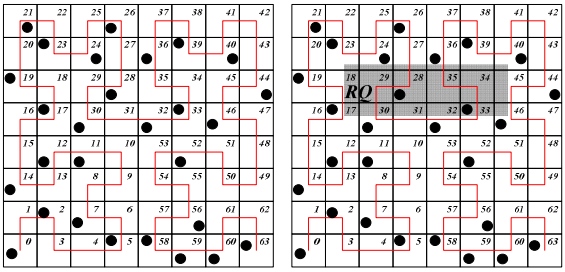
James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_7,

commonly used in the wireless broadcasting system: *Access Latency* and *Tuning Time*. The former is the duration elapsed from the moment a MC receives the query from its user to the moment the query is satisfied. The latter is the duration during which the MC remains in the full operational mode, called the *active mode*, which is proportional to the amount of the energy consumed by the MC. Both *Access Latency* and *Tuning Time* are evaluated in terms of the number of buckets (the smallest logical unit of wireless data broadcast). In order to support efficient query processing in wireless data broadcast, the air index scheme has to minimize *Tuning Time* while minimizing *Access Latency* by keeping the increase of the *broadcast cycle* minimal.

In this paper, we propose a novel index structure, called Distributed Space-Partitioning Index (DSPI), for supporting range queries in wireless data broadcast environments. DSPI consists of a hierarchy of grids, each of which is constructed by partitioning the data space. Additionally, DSPI has a linear, and yet distributed structure suitable for wireless data broadcast. The rest of the paper is organized as follows. In Section 2, related work to our study is presented. Section 3 presents the proposed index scheme, namely DSPI. In Section 4, experiments are conducted to show the efficiency of DSPI. Finally, we conclude this paper in Section 4.

2 Related Work

Recently, several index structures for supporting LDQs in wireless data broadcast have been proposed. In [3], a linear index structure, referred to as HCI, is proposed based on the *Hilbert curve* (HC). In HCI, each LDD is assigned a unique HC value by recursively partitioning the data space into $2^{\lambda \cdot 2}$ cells, until each cell contains only one LDD, where λ is the order of HC. HC needs to allocate a sufficient number of bits to represent the HC values in order to guarantee that each LDD in the original space is assigned a unique HC value. Figure 1a illustrates an HC($\lambda = 3$) and a sample data set. By using HC values, HCI constructs B^+-tree, where each HC value of the LDD acts as the key value.



(a) HC(λ = 3) (b) Range query

Fig. 1. HC($\lambda = 3$) and HC based range query processing

DSI, presented in [4], also adopts HC for index organization. In order to organize the index in a distributed fashion, DSI divides the whole set of LDD into n_F frames, each of which has the index table which maintains HC values of $\log_r n_F$ frames and pointers to the frames, where r is the exponential base, and n_F is the number of frames in a *broadcast cycle*.

In both HCI and DSI, MCs have to listen to the broadcast channel more than necessary because the index is constructed based on the HC value instead of the coordinates of the LDD. As illustrated in Figure 1b, in order to process range query RQ (shaded area), an MC has to read the LDD whose HC value $\in$ {17, 23, 24, 28, 30, 32, 33, 36, 39, 40, 46}. However, only the LDD, whose HC value $\in$ {28, 33}, can be the result of *RQ*. This leads the MC to receive unnecessary LDD and thus, increases *Tuning Time* of the MC. Furthermore, not only the distribution of LDD in the original space but also the number of LDD may greatly deviate the performance of both HCI and DSI in terms of *Access Latency* as well as *Tuning Time*.

3 Distributed Space-Partitioning Index (DSPI)

In this section, we propose Distributed Space-Partitioning Index (DSPI) that enables MCs to selectively receive the required data instances (e.g., LDD), and thus reducing *Tuning Time* performance, with shorter *Access Latency* than that of previous indices [3, 4]. DSPI consists of n_G hierarchical levels of grids, each of which takes different granularity from the others, where n_G is the number of grids. Figure 2 shows the basic idea of DSPI, where $n_G = 2$. Without loss of generality and for simplicity, we assume that $n_G = 2$ in the rest of the paper. However, extension to $n_G > 2$ is straightforward. The lowest level of the grids, referred to as *Leaf Grid*, points to the data instances, whereas, the upper level of the grids, called *Directory Grid*, points to the *Leaf Grid*. For the linear placement of the data on the air, *Hilbert curve* (HC) values are assigned to each cell as an identifier, according to its occurring order on the HC (See Figure 2). It is known that HC is effective in mapping 2-dimensional space to 1-dimensional linear space while preserving the spatial locality of the underlying data.

Definition 1. *(Leaf Grid). The data space DS is a two-dimensional space. The Leaf Grid uniformly partitions the DS into* $2^{\gamma_L \cdot 2}$ *cells (leaf cells) of* $= \frac{1}{2^{\gamma L}} \times \frac{1}{2^{\gamma L}}$ *, where* γ_L (≥ 1) *is the granularity factor of the Leaf Grid. Leaf Grid, which takes the finest granularity, provides MCs with local view of the entire database D.*

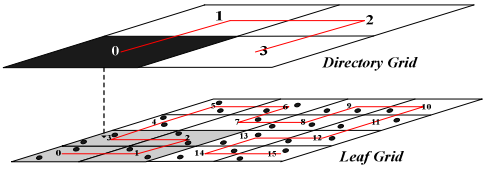
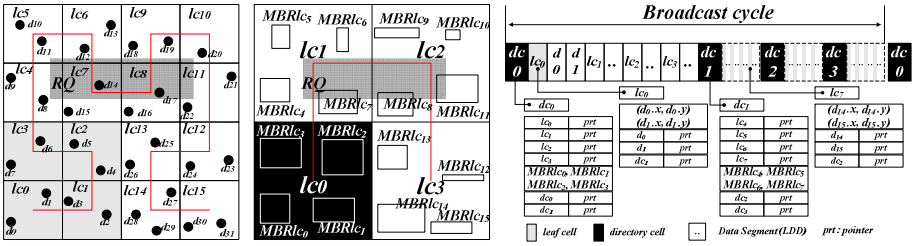


Fig. 2. Hierarchical grids ($n_G = 2$)

Definition 2. *(Directory Grid). The Di1rectory Grid uniformly partitions the DS into* $2^{\gamma D \cdot 2}$ *cells (directory cells) of size* $= \frac{1}{2^{\gamma D}} \times \frac{1}{2^{\gamma D}}$, *where* $1 \leq \gamma_D \leq \gamma_L$ *and* γ_D *is the granularity factor of the Directory Grid. In particular, the Directory Grid which takes the coarsest granularity is called the Root Grid. Since we assume* $n_G = 2$, *hereafter we use Directory Grid and Root Grid interchangeably. Directory Grid maintains information on the lower level grid, namely, Leaf Grid* ($\because$ $n_G = 2$) *as well as the global view of D.*



(a) *Leaf Grid* (b) *Directory Grid* (c) Broadcast channel structure

Fig. 3. Organization of the hierarchical grids

Both *Leaf Grid* and *Directory Grid* constitute the hierarchical grids. Figure 3a, 3b and 3c illustrate the *Leaf Grid*, the *Directory Grid* and the broadcast channel structure respectively. A *leaf cell* lc_{id}, contained in the *directory cell* dc_{id}, is the *child cell* of the dc_{id}. For example, *lc*0 ~ *lc*3 in Figure 3a are *child cells* of the dc0 in Figure 3b. Specifically, each *leaf cell* of the Leaf Grid maintains the following information:

- Identifier, dc_{id} (being assigned according to the occurring order on the HC of order $\lambda = \gamma_L$).
- Coordinates of the data instances contained in it.
- Pointers (arrival time) to the data instances contained in it.
- Pointer to the *next directory* cell; for example, lc_0 ~ lc_3 include the pointer to the dc_1.

Note that *leaf cells* are sorted according to their identifiers, and each leaf cell is placed in front of the data segment pointed by it on the broadcast stream. Each *directory cell* of the *Directory Grid includes* the following information:

- Identifier, dc_{id} (being assigned according to the occurring order on the HC of order $\lambda = \gamma_D$).
- Pointers to the *child cells*.
- Minimum bounding rectangles, each of which, referred to as $MBRlc_{id}$, completely encloses all the data instances contained in the corresponding *child cell* lc_{id} (See Figure 3b).
- $\log_2 N$ pointers p_i, each of which points to the 2^i*th directory cell*, where $0 \leq i \leq \log2(N) - 1$ and N is the number of *directory cells*. For example, dc_0 has the pointers to dc_1 (the next 2^0*th directory cell*) and dc_2 (the next 2^1*th directory cell*).

As shown in Figure 3c, *directory cells* are sorted according to their identifiers and each of them is placed in front of its *child cells* on the broadcast stream. A range query *RQ* retrieves all data instances within a given query rectangle.

A range query *RQ* retrieves all data instances within a given query rectangle. To process an *RQ* with DSPI, an MC determines a collection of *leaf cells*, which overlap with the query rectangle, and thereafter it retrieves the qualified data instances by traversing the *directory cells* and *leaf cells*. For example, consider the range query *RQ* (shaded area) in Figure 3a and 3b. As a first step, the MC determines a set of overlapped *leaf cells* ($lc_4 \sim lc_{11}$), referred to as OC. Then, the MC tunes into the *directory cell* dc_{id}, which is the next nearest *directory cell* from the current position. Without loss of generality, we assume that the MC first listens to the broadcast channel at the time (or position) of dc_1, which contains the information of its *child cells* ($lc_4 \sim lc_7$) as well as $MBRlc_4 \sim MBRlc_7$. Because the $MBRlc_7$ overlaps with the *RQ* as shown in Figure 3b, the MC reads lc_7 and selectively retrieves the qualified data instance namely, d_{14} by utilizing the coordinates maintained in lc_7. The MC removes $lc_4 \sim lc_7$ from the OC and moves to the next *directory cell* dc_2 by using the pointer maintained in either dc_1 or lc_6. Since $MBRlc_8$ and $MBRlc_{11}$ overlap with the *RQ* (See Figure 3b), the MC reads lc_8 as well as lc_{11} and retrieves d_{17}. In this way, the MC completes *RQ* after visiting dc_1 and dc_0. Algorithm presents the detailed algorithm for processing range queries based on DSPI.

Algorithm. Range Query Processing

Input RQ: Range Query,
OC: a set of the overlapped *leaf cells*
Output *QueryResult*: data instances within the *RQ*
Procedure
1: listen to the broadcast channel
2: **do** {
3: move to the next *directory cell* (*root cell*)
4: **for** each *child cell* (*leaf cell*) $\in$ *OC* **do**
5: **if** ($MBRlc_{id}$ overlap with *RQ*) **then**
6: **for** each data instance d_{id} in the *child cell* **do**
7: **if** (d_{id} is within the *RQ*) **then**
8: result = result $\cup$ d_{id}
9: **end if**
10: **end for**
11: **end if**
12: remove *child cell* from the *OC*
13: **end for**
14: }**while** (*OC* is not empty)
15: **return** result

Fig. 4. Pseudo code of Range Query Processing algorithm

4 Performance Evaluation

We evaluated the performance of DSPI by comparing its *Access Latency* and *Tuning Time* with those of HCI and DSI. For the implementation of the HCI, the (1, *m*) indexing technique was utilized, where the global index is inserted every $\frac{1}{m}$ fraction of the *broadcast cycle*, and optimal m (= $\sqrt{\frac{DATA_{size}}{INDEX_{size}}}$)was used in order to reduce *Access Latency*, where $DATA_{size}$ and $INDEX_{size}$ are the sizes of whole data instances and the global index in terms of the number of buckets [1]. On the other hand, for the DSI, the exponential base *r* was set to 2 and the number of data instances within each frame was determined so that index table associated with the frame could fit into one bucket [4]. For the configuration of DSPI, we set $n_G = 2$ for simplicity. In addition, the granularity factor, γ_L of the *Leaf Grid* was set to 4, and the corresponding granularity factor of *Directory Grid* was set to 3. All the experiments were conducted on a Pentium IV 3.2 GHz machine with 1 GB RAM. The system model, which consists of a base station, MCs and a broadcast channel, was implemented in the Java language. We conducted performance analysis on two datasets: UNIFORM dataset and REAL dataset. In the UNIFORM dataset, 6,000 data instances are uniformly generated, while the REAL dataset contains 5,848 actual cities and villages of Greece [5]. In the experiment, 10,000 queries were issued. We use the number of bytes instead of the number of buckets to evaluate *Access Latency* and *Tuning Time* because the bucket size is varied in the experiment [4]. The number of bytes can be translated into the time values without loss of generality, since the bandwidth of wireless channel is fixed. The size of a data instance is 1024 bytes, and 16 bytes are used for representing the 2-dimensional coordinates of a data instance (8 bytes each) as well as the HC value (16 bytes). We set the pointer size to 2bytes. Table 1 illustrates parameter settings used in the experiments.

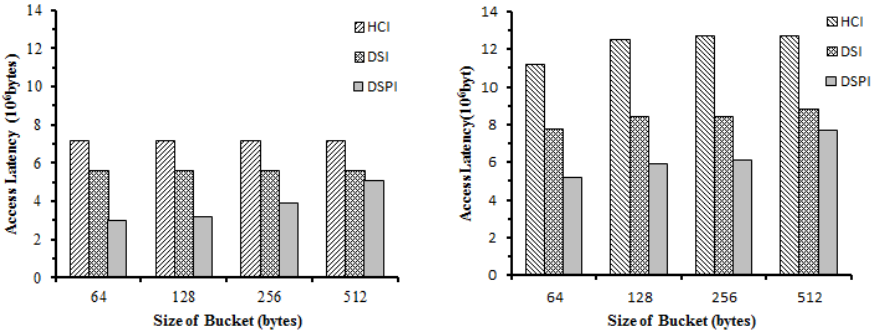
Table 1. Parameter Settings

Parameter	Setting
bucket size	2^6 ~2^9 bytes
LDD size	1204 bytes
coordinates size	16 bytes
pointer size	2 bytes

Since the number of MCs does not affect the system performance [3], we considered only the performance aspects of an MC.

Evaluation of Access Latency. Figure 5 shows the *Access Latency* performance. For both datasets, namely UNIFORM (Figure 5a) and REAL (Figure 5b), DSPI is superior to the other two indexing schemes (HCI, DSI). This is due to the fact that the overall length of the *broadcast cycle* in DSPI is shorter than those of the other two indices. In other words, the index size of HCI and DSI is much larger than that of DSPI. In HCI, the whole index is interleaved *m* times with data instances, while in

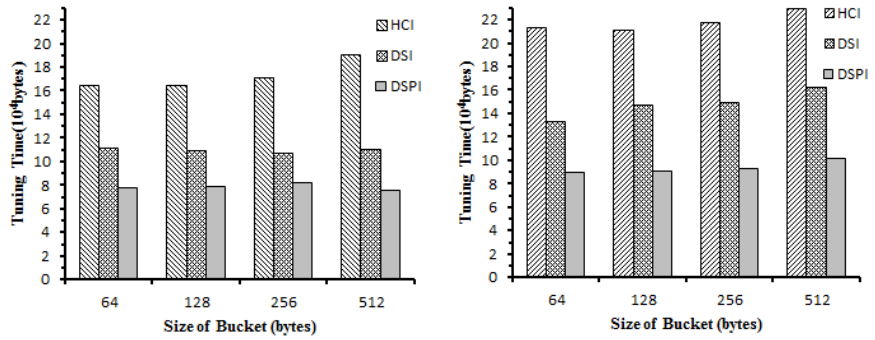
DSI, every frame has to maintain lots of information (e.g., HC values of $\log_r n_F$ frames and pointers to those frames) when a large number of data instances are concerned. Furthermore, in DSI, redundant information increases according to the increase of the number of data instances.



(a) UNIFORM dataset (b) REAL dataset

Fig. 5. Comparison of *Access Latency*

Evaluation of Tuning Time. Note that the *Tuning Time* affects the energy consumption on MCs. Figure 6 denotes the *Tuning Time* performance of the three schemes. As shown in the figure, DSPI outperforms HCI and DSI. This is because DSPI allows MCs to avoid reading unnecessary data instances by offering the global view as well as the local view. As mentioned in Section 2, both HCI and DSI are constructed based on the HC value instead of the coordinates of the underlying data instances. As a result, MCs have to listen to broadcast channel more than necessary. This incurs a huge amount of *Tuning Time*. Furthermore, as the order of HC increases (due to the skewed distribution or a large number of data instances), spatial locality of the data instances in 2-dimensional space can not be preserved in the 1-dimensional linear space. Therefore, MCs read much more data instances than necessary.



(a) UNIFORM dataset (b) REAL dataset

Fig. 6. Comparison of *Tuning Time*

5 Conclusion

This paper proposed a novel index structure, called Distributed Space Partitioning Index for processing range queries in wireless broadcast environments. DSIP utilizes the notion of hierarchical grids in order to allow MCs to selectively retrieve the required data in an efficient way in terms of *Access Latency* as well as *Tuning Time*. As demonstrated in the simulation study, DSPI outperforms existing index schemes (e.g., HCI, DSI). As future work, we plan to investigate an efficient algorithm for supporting k-nearest neighbor (k-NN) queries based on DSPI.

References

1. Imielinski, T., Viswanathan, S., Bardrinath, B.R.: Data on air: Organization and access. IEEE Trans. TKDE 9(3), 353–372 (1997)
2. Xu, J., Lee, W.C., Tang, X., Gao, Q., Li, S.: Exponential index: a parameterized distributed indexing scheme for data on air. In: Proc. 2nd ACM Conf. MobiSys 2004, Boston, Massachusetts, USA, pp. 153–164 (June 2004)
3. Zheng, B., Lee, W., Lee, D.: Spatial queries in wireless broadcast systems. Wireless Network 10(6), 723–736 (2004)
4. Lee, W., Zheng, B.: DSI: A Fully Distributed Spatial Index for Location-based Wireless Broadcast Services. In: Proc. 25th IEEE Conf. ICDCS 2005, Columbus, Ohio, USA, pp. 349–358 (June 2005)
5. Spatial Datasets, `http://www.rtreeportal.org.sp-atial.html`

End-to-End High Speed Forward Error Correction Using Graphics Processing Units

Md Shohidul Islam and Jong-Myon Kim*

School of Electrical Engineering, University of Ulsan, 93 Daehak –ro, Nam-gu, Ulsan 680-749, Korea
shohid@mail.ulsan.ac.kr, jmkim07@ulsan.ac.kr

Abstract. Forward error correction (FEC) is an efficient error recovery mechanism for wireless networks in which erroneous packet is corrected in the destination node. More importantly, real-time and high-speed wireless networks require fast error recovery to ensure quality of service (QoS). Since graphics processing units (GPUs) offer massively parallel computing platform, we propose a GPU-based parallel error control mechanism using extended Hamming code supporting single-bit as well as multiple-bit error correction. We compare the performance of the proposed GPU-based approach with the equivalent sequential algorithm that runs on the traditional CPU for error strength, t, such that $1 \leq t \leq 7$. Experimental results demonstrate that the proposed GPU-based approach outperforms the sequential approach in terms of execution time. Moreover, the proposed parallel implementation yields significant reduction in computational complexity from $\boldsymbol{O}(n^3)$ of the sequential algorithm to $\boldsymbol{O}(n)$ of the GPU-based approach, leading to tremendous speedup gain.

Keywords: Real-time wireless communication, multiple bit error FEC, extended Hamming code, GPU.

1 Introduction

Wireless communications unleash promising arena with the diversity from wireless internet access, telecommunication, voice over internet protocol (VOIP), wireless video multicasting, mobile adhoc network (MANET), wireless sensor network (WSN) and so on. However, signals in wireless medium are highly prone to fading, interference or noise that results in erroneous data. Inherent vulnerability to packet corruption imposes paramount challenge for reliable communication [1]. Such error is another cause of packet loss at destination. As a result, error control is necessary to guarantee fair and consistent data delivery. In communication systems, FEC is one of the effective error recovery mechanisms in which original packets are encoded with some redundant information in the transmitter before transmitting over wireless medium[1], [2]. Upon receiving in the destination end, packets are decoded using the same redundancy by some FEC coding algorithms for possible error detection and correction.

Error correction codes are popularly being employed in long distant information transferring or channel where message might get corrupted and it becomes performance

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_8,

bottleneck [3]. Fundamentally error codes can be either Convolution codes or Block codes where the former one possesses bit by bit processing and suitable for hardware implementation. The later codes are processed in blocks and suitable for software implementation. Suitable techniques are preferred depending on a number of factors including error pattern which is the most crucial. Errors can fall in random, single bit, or multiple bit errors. High reliability areas require adequate protection that single error correction (SEC) may not provide. Under these circumstances, multiple error correction (MEC) is highly desirable [3]. More importantly, real-time, high speed and other time sensitive wireless networks demand fast error recovery since massive data is generated and need to be processed momentarily.

Inspired by the massively parallel computing capability, low cost but extensive programmability and scalability, GPU is recently being exploited in general purpose beside graphics rendering [5-7]. Considering these factors, in this paper, we propose GPU based faster error recovery using extended Hamming code [2] that can support single bit as well as multiple bit errors in the packet level [4]. We validate our proposed approach using a compute unified device architecture (CUDA) [5] enabled NVIDIA GeForce GTX 560 graphics card. We compare the performance of parallel FEC algorithm with the equivalent sequential algorithm that run on the traditional CPU. The experimental results demonstrate that the GPU-based approach outperforms the serial approach in terms of execution time. Furthermore, the proposed implementation yields significant reduction in computational complexity from $\boldsymbol{O}(n^3)$ of the sequential algorithm using CPU to $\boldsymbol{O}(n)$ of the GPU-based approach, achieving tremendous speedup.

The rest of the paper is organized as follows. Section 2 describes the implementation of extended Hamming code as FEC, Section 3 presents experimental results and analysis. Finally, Section 4 concludes the paper.

2 FEC Implementation

We implement extended Hamming coding for multiple bit error detection and correction on CPU and GPU. This section briefly introduces the fundamental principle of the algorithm and its implementation.

2.1 Overview of Extended Hamming Coding

Describing the mechanism of entire algorithm requires understanding the following definitions.

- *Codeword*: An ordered set (h, m, r) is called Hamming codeword which refers m bit information is appended by r redundant bits to produce encoded packet of h bits. Thus, code length, h, is given by $h=m+r$ and r is determined by the inequality such as $2^r \geq m+r+1$.
- *Code rate*: The ratio of total information bits (m) to code word size (h) is said to be Hamming code rate.

- *Error strength (t):* For the maximum number of corrupted bits in a data packet, the algorithm is capable to detect and correct. It is determined by the channel condition and remains available in both sender and receiver end.

Extended Hamming coding utilizes fundamental Hamming coding. A typical (11, 7, 4) coding has *7-bit* encoded segment, *H*, from *7-bit* main message, *M*, by adding *4-bit* redundancy following the linear operation:

$$H=M \bullet G \tag{1}$$

where *G* is the generator matrix of the (11, 7, 4) Hamming coding. Extended Hamming coding on the sender side consists of splitter, encoder and merger, as shown in Fig. 1, in which a packet of *M* bit is fragmented into $M_1, M_2, \cdots, M_t$ segments. Encoder employs fundamental Hamming coding of (*h*, *m*, *r*) on each segment resulting in encoded sequence $M_1^{'}, M_2^{'}, \cdots, M_t^{'}$. Finally, merger unifies all these incoming sequences to generate consolidated Hamming coded packet $M^{'}$, which is transferred as a radio signal over wireless medium by the transmitter.

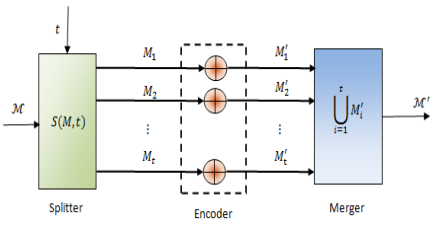


Fig. 1. Stage at the sender end

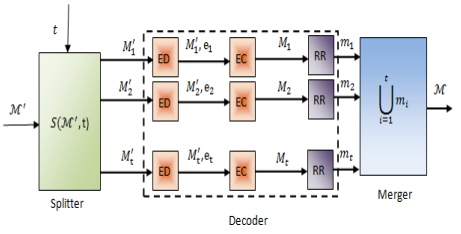


Fig. 2. Stage at the destination end

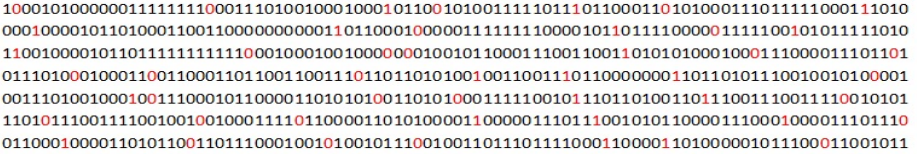


Fig. 3. Error positions in the packet detected at destination

Refer to Fig.2, the algorithm at the destination end accomplishes segmentation of the received packet, $M^{'}$. Each individual segment is attempted in the error detection (ED) stage in order to determine possible error if takes place. Thereafter, error correction (EC) unit flips the bits in the positions reported by ED and transferred to the redundancy removing (RR) phase. Redundant information is not a part of the original message. They are exclusively meant for error detection, which is why removing them is imperative. In the end, all the fragments m_1, m_2,…,m_t are

concatenated to obtain the source main message, *M*. Entire stages in Fig. 2 are implemented on CPU and GPU. For the time sensitive wireless networks, fast error recovery at the destination is highly desirable. Fig. 3 and Fig. 4 show a part of the sample experimental output for packet size *M=64* and *t=7*, respectively. However, it can support variable length packet size and larger error strength as well if required dynamically.

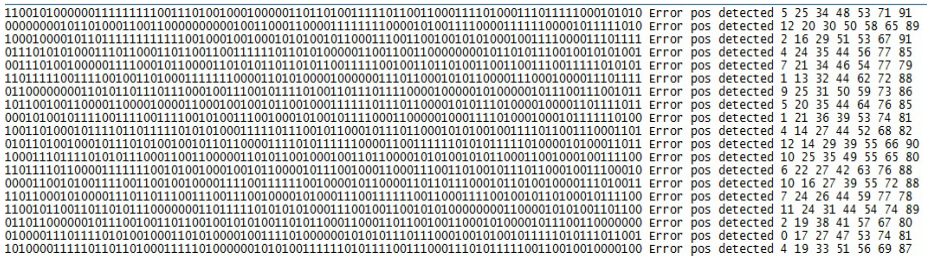


Fig. 4. Error correction at destination

2.2 Implementation Mechanism

Three steps such as Splitter, Decoder and Merger of FEC are implemented in two different ways: the sequential approach is run on CPU which is quite straightforward as explained in Section 2.1 and the equivalent parallel approach is implemented on GPU.

Table 1. Experimental Environment

Property	CPU	Property	GPU
Processor	Intel(R) Core(TM) i5-3570K	Brand name	NVIDIA GeForce GTX 560
Clock speed	3.40GHz	Processor clock	1620MHz
No of Cores	4	CUDA core	336
No of threads	4	Total MP	7
RAM	8.00GB	Max thread per block	1024
Bus/Core ratio	34	Shared memory per MP	49152 Byte
Operating system	Win 7, 32 bit	Total global memory	1 GB

The system specification is summarized in Table 1. Regarding to GPU implementation, one block is assigned to compute one checksum bit of each individual segment, $M_1^{'}, M_2^{'}, \cdots, M_t^{'}$, in parallel. Therefore, the total number of

blocks required to be declared is $\sum_{i=1}^{t}\sum_{j=1}^{R(M_i^{'})} r_j$ where $R(M_i^{'})$ is the total number of redundancy for segment $M_i^{'}$ and r_j is its j-th redundant bit respectively.

3 Experimental Results

This section presents the performance of CPU and GPU for the multiple bit error correction in terms of execution time.

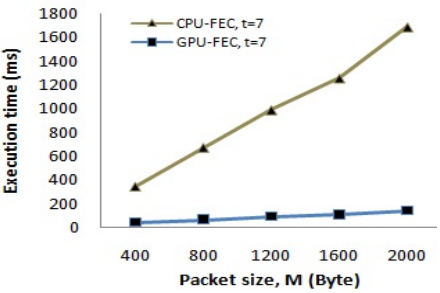


Fig. 5. Execution time vs. packet size

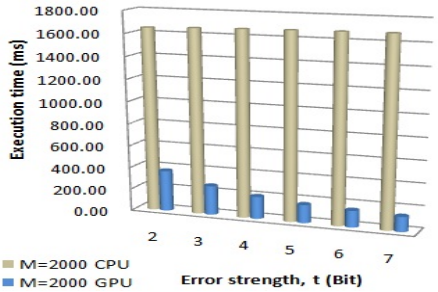


Fig. 6. Execution time vs. error strength

Table 2. Execution time (ms) on CPU

Packet size, M (Byte)	t=2	t=3	t=4	t=5	t=6	t=7
400	298.73	305.86	309.80	319.00	322.94	339.42
800	628.11	641.14	648.52	654.31	659.64	667.29
1200	929.58	939.24	943.35	951.47	958.64	983.94
1600	1202.10	1216.53	1224.72	1235.58	1244.81	1251.61
2000	1648.20	1659.95	1667.49	1674.19	1679.28	1682.67

Table 3. Execution time (ms) on GPU

Packet size, M (Byte)	t=2	t=3	t=4	t=5	t=6	t=7
400	89.30	66.16	52.03	44.91	41.47	39.63
800	165.67	116.84	92.38	76.88	69.54	60.99
1200	240.99	163.72	128.90	106.18	96.35	88.19
1600	312.05	214.87	163.70	139.50	119.43	103.63
2000	370.48	264.14	202.04	164.67	150.91	135.68

Tables 2 and Table 3 show the execution time of sequential and parallel FEC in milliseconds using CPU and GPU, respectively, for various packet size, M, and error strength, *t*. For a fixed size of packet, CPU runtime reaches to large extents and linearly increases with the increase of bit errors in the packets. In contrast, GPU finishes the equivalent concurrent processing with much lower time, resulting in tremendous speedup. Fig. 6 is an instance showing such trend for packet size of 2000 Bytes. Refer to Fig. 5, exponential rising tendency of CPU time is reduced to linear order by GPU when packet size is varied and error strength remains the same. The fundamental underlying reason is that an increase in the packet size or the error strength increases the Hamming code length and amount of checksums to calculate. For GPU-based error correction, a block is assigned to generate a checksum, and numerous similar threads and blocks end up these computations simultaneously on streaming multiprocessors. In contrast, sequential pattern of checksum calculation on CPU in essence causes long latency.

4 Conclusions

In this paper, we proposed a novel GPU based fast forward error correction (FEC) approach using extended Hamming coding. This end-to-end packet level FEC can provide both single bit as well as multiple bit error resiliency to the network, boosting reliability in communication. Moreover, the algorithm can dynamically adapt various code lengths and error strengths depending on channel condition. Quality of service in time sensitive networks is heavily bounded by the packet processing latency. Experimental results demonstrated that the proposed GPU based FEC approach extremely outperforms the CPU based sequential approach in terms of execution time and computational complexity. Consequently, the approach can be effectively applied in time sensitive and high speed wireless communication systems.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566), and by the Leading Industry Development for Economic Region (LeadER) grant funded by the MOTIE(The Ministry of Trade, Industry and Energy), Korea in 2013. (No. R0001220).

References

[1] Tsai, M., Shieh, C., Huang, T., Deng, D.: Forward-Looking Forward Error Correction Mechanism for Video Streaming Over Wireless Networks. IEEE Systems Journal 5(4), 460–473 (2011)

[2] Xu, J., Li, K., Min, G.: Reliable and Energy-Efficient Multipath Communications in Underwater Sensor Networks. IEEE Transactions on Parallel and Distributed Systems 23(7), 1326–1335 (2012)

[3] Singh, J., Singh, J.: A Comparative Study of Error Detection and Correction Coding Techniques. In: Proc. 2012 Second International Conference on Advanced Computing & Communication Technologies (ACCT), January 7-8, pp. 187–189 (2012)

[4] Hund, J., Heinrich, A., Ziller, A., Schwingenschlogl, C., Kraemer, R.: A packet-level adaptive forward error correction scheme for wireless networks. In: 2010 7th Workshop on Positioning Navigation and Communication (WPNC), March 11-12, pp. 1–3 (2010)

[5] Sanders, J., Kandrot, E.: CUDA by Example: An Introduction to General-Purpose GPU Programming, 1st edn. (July 29, 2010), http://www.amazon.com/CUDA-Example-Introduction-General-Purpose-Programming/dp/0131387685

[6] Kirk, D.B., Hwu, W.W.: Programming Massively Parallel Processors: A Hands-on Approach, 2nd edn. (December 28, 2012), http://www.amazon.com/Programming-Massively-Parallel-Processors-Edition/dp/0124159923/ref=dp_ob_title_bk

[7] Bilel, B.R., Navid, N.: Cunetsim: A GPU based simulation testbed for large scale mobile networks. In: Proc. 2012 IEEE International Conference on Communications and Information Technology, June 26-28, pp. 374–378 (2012)

DirectSpace: A Collaborative Framework for Supporting Group Workspaces over Wi-Fi Direct

Jong-Eun Park, Jongmoon Park, and Myung-Joon Lee*

School of Electrical Engineering, University of Ulsan, Korea
cjswowhddms@nate.com, monster28g@gmail.com,
mjlee@ulsan.ac.kr

Abstract. Wi-Fi Direct is a new feature in the 4.0 version of the Android operating system, allowing devices to connect directly to each other via Wi-Fi without an intermediate access point. Based on the Wi-Fi Direct feature, in this paper, we present a collaborative framework named DirectSpace that supports various collaborative situations to resolve the problems of traditional application paradigm such as client-server model. For this, we design collaborative services to support various environments, and develop collaborative protocols for providing workspaces over Wi-Fi Direct. On the top of DirectSpace, we also present a collaborative application for utilizing the group workspaces.

Keywords: Collaborative framework, Wi-Fi Direct, DirectSpace, workspace.

1 Introduction

In 2012, Google has released a version of the android operating system 4.0(Ice Cream Sandwich)[1], where the new Wi-Fi Direct feature allows devices to connect directly to each other via Wi-Fi network without an intermediate access point. So, regardless of the availability of mobile networks, it can support a useful communication environment for smart devices to share any information such as video, photos, files and documents.[2,3]

In this paper, we describe the development of a collaborative framework named DirectSpace that supports group workspace among multiple devices via Wi-Fi Direct. This framework provides an environment to effectively share collaborative workspaces with nearby devices. The collaborative workspaces support access control mechanism and group management[4] facility for effective collaboration. The collaborative services provide instant messaging and resource sharing based on the protocols designed for the framework. In addition, we present a collaborative application running on Android through the API provided to effectively use the DirectSpace, which leads to easy development of collaborative application.

This paper is organized as follows. In section 2, we describe the design of collaborative services, workspaces and collaboration protocols, and the development

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_9,

of DirectSpace. Section 3 explains the implementation of the collaborative application running on Android. In section 4, we discuss different collaborative systems with similar features and section 5 presents some concluding remarks.

2 DirectSpace over Wi-Fi Direct

In this section, we describe the design of collaborative services based on Wi-Fi Direct and present the implementation of DirectSpace utilizing the services.

2.1 Design of Collaborative Services and Workspaces

In DirectSpace, each peer acts as both a server and a client at the same time. When a peer acts as a server, the server peer provides a workspace for collaboration; otherwise, the peer acts as a client to the workspace provided by some server peer. Table 1 shows the collaborative services defined in DirectSpace.

Table 1. Collaborative Services in DirectSpace

Service Type	Service List
Connection Service	-Peer Discovery Service, -Connection/Disconnection Service
Collaborative Service	-Resource Sharing Service -Group Managing Service -Instant Messaging Service

The connection service is basically provided by Wi-Fi Direct, being used to discover other peers and connect to them. According to the Wi-Fi Direct specification, once two or more devices connect directly, they have formed a Wi-Fi Direct group. All Wi-Fi Direct devices must be able to negotiate which device adopts the ‘group owner’ role when forming a group. The group owner has an ‘AP-like’ capability, controlling the group and enabling connectivity.

After peer discovery is completed successfully, every peer can connect each other and form a group in a transmission range of Wi-Fi Direct. The resource sharing service provides facilities like as the file explorer to manage and share its own resources and others in the SD Card (Secure Digital Card). The group managing service supports the configuration of various user-defined group workspaces and the associated access control for effective collaboration. The messaging service is used to transmit both instant messages and any information related to collaboration asynchronously to group users. Users can share resources through the workspace managed as a folder on the external storage by the server peer. Figure 1 shows the workspace structure in the Android file system. Each folder in depth level 1 is treated as a workspace while the folder depth level greater than 1 is considered as collaborative folders and files.

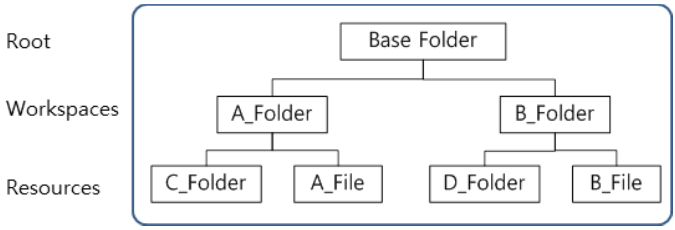


Fig. 1. Hierarchical Folder-based Workspace Structure

Since a mobile peer-to-peer network via the Wi-Fi Direct feature is frequently reconfigured, resource damage and loss should be prevented through the appropriate authorization and resource access control in a workspace. A general workspace allows all users to access, whereas a group workspace allows only authorized users to access. Both types of workspaces can be protected through the passwords specified by the server peer. To support the workspaces of these types, the service checks the state of each connected user, managing the user according to the states in the workspace. The state of a peer is classified into the following 3 ones: One is the *naive* state that a peer does not join the workspace yet. Another is the *waiting* state after making a request to join a group workspace and the last one is the *joined* state to the group workspace after approval of the server peer. The access level is classified into download-only, upload only and full access.

2.2 Design of Collaborative Protocols

A peer connected via Wi-Fi Direct can act as a server and a client at the same time. In other words, each peer can respond to a request while making a request simultaneously. The collaborative protocol is designed to request and respond for cooperation of a peer group. The protocol consists of arrays of key-value pair as a JSON type:

```
{"Request": a type of collaborative service, "Workspace":
workspace to be cooperated, "Sender": information of
request peer, "Detail": new JSON Object (an additional
specific condition)}
```

The attribute "Sender" which contains information of the request peer is used to return results of the request. The attribute "Detail" represents the detailed information for handling the request.

2.3 Implementation of DirectSpace

A workspace is instantiated through a Java class that abstracts a workspace in physical SD Card of a device. The class is represented by a type of workspace, access permissions, and an organized peer list. Since instances of the workspace class are

destroyed when the framework is terminated, an instance of the workspace class is stored into the SD Card for effective and safe management as follows:

```
"Workspace": JSONObject ("Name": title of workspace
"Authority": permission of workspace, "Open": public
access (yes/no), "Group": limited access (yes/no),
"Password": secret key
"Group member": new JSONObject ("Name": display of peer,
"Status": status of peer))
```

Unfortunately, the workspace created in SD Card can be removed or modified by another application because it is considered as a normal folder by the android system. In this case, a fault of workspace synchronization occurs between the logical state of the workspace and the physical state in the SD Card. To prevent this problem, it is necessary to verify whether the current workspace is exactly same as the saved workspace whenever the configuration file for workspaces is processed. After the configuration file is processed, workspaces should be checked periodically while the application is running. Figure 2 shows how it works.

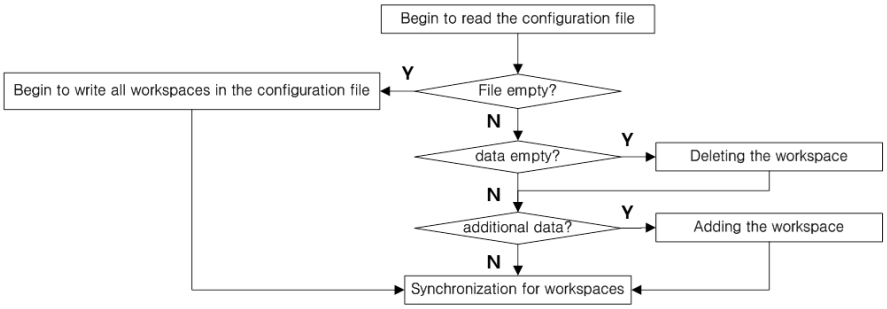


Fig. 2. Procedure for Workspace Synchronization

The collaborative services consist of a communication module and a protocol parser. The communication module provides functions to perform both of server and client roles. The protocol parser analyzes the transmitted protocol messages and sends results to the collaborative services. DirectSpace provides collaboration API to develop collaboration applications in a systematic manner over Wi-Fi Direct. Figure 3 illustrates the structure of the DirectSpace collaborative framework. In a typical Wi-Fi Direct environment, an ownership is determined after a network topology is established. Then, every user can communicate with each other via the group owner. In general, IP address is invisible except group owner over the Wi-Fi Direct network, whereas MAC(Media Access Control) address is available. To support the IP network topology, DirectSpace uses ARP(Address Resolution Protocol) to translate MAC addresses into IP addresses. After address translation, every peer is able to directly communicate with each other without the need for the group owner, which reduces the communication overhead for redirection through the group owner.

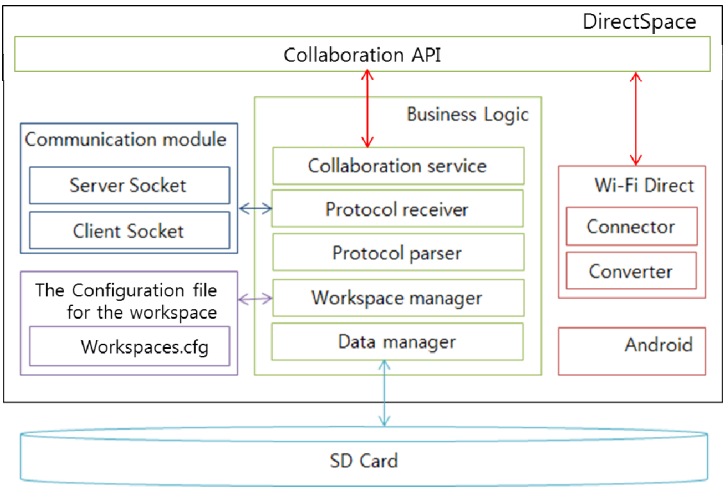


Fig. 3. Structure of the DirectSpace collaborative framework

3 Application of DirectSpace

With the help of API provided by DirectSpace, we develop a collaboration application on the android system. The application supports a real-time instant messaging, resources sharing and connecting facilities via Wi-Fi Direct. The application also supports access control mechanism for workspaces in various collaborative situations. A server peer configures the permission rules and types of its workspaces, managing participants in its workspaces. A client peer can connect to a specific workspace and access resources according to the designated authority level. Figure 4 shows the application interface for collaborative services.

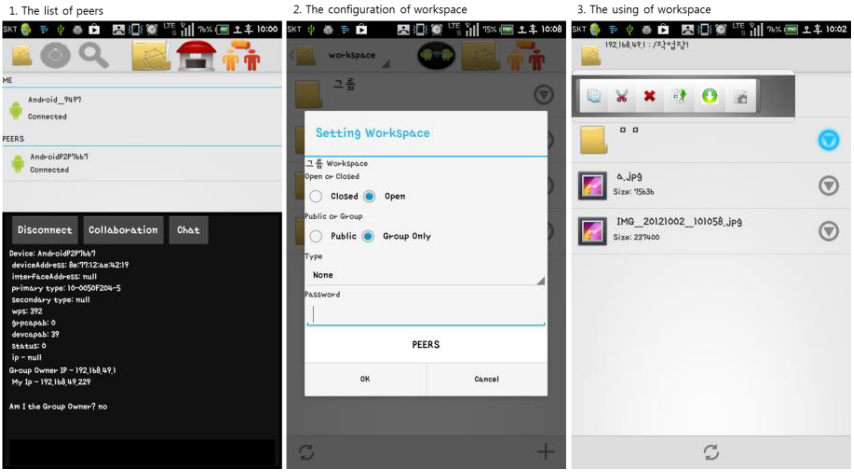


Fig. 4. User interface of the Collaboration Application using DirectSpace

4 Comparison with Related Works

As smart devices proliferate in the mobile network environment, there arise a variety of situations that require cooperation. WareBox[5] is a typical collaborative application that consists of a server and a client based on the WareBox development platform. The platform relies on a social network service and collaborative services. WareBox supports reusability and productivity by combining the provided collaboration features. SuperBeam[6] is an application providing a way to share large files between Android devices using QR codes or NFC through Wi-Fi Direct. Table 2 shows a comparison with these applications.

Table 2. Functional Comparison with Other Systems

Features	WareBox	SuperBeam	DirectSpace
Communication paradigm	Client/Server	Wi-Fi Direct	Wi-Fi Direct
Access point	O	X	X
Group management	O	X	O
Upload, Download	O	O	O
General workspaces	O	X	O
Group workspaces	O	X	O
Password-protected feature	X	X	O

WareBox provides collaborative services based on various workspaces, supporting group messaging service using Google push service. Since WareBox relies on the client-server model, it is vulnerable to the network failure or the server failure. Also, it does not support the ad-hoc network environment.

Based on Wi-Fi Direct that supports peer-to-peer network, SuperBeam and DirectSpace provide file sharing service on smart devices without any infrastructure network. Since SuperBeam does not support the workspaces and group management service, it cannot meet the collaboration requirement arisen on various situations. However, DirectSpace can handle those various situations by providing group workspaces for collaborative services on smart devices. This feature immediately provides a fully distributed environment to spontaneously share several types of workspaces with nearby devices because it does not rely on central infrastructure.

5 Conclusion

In this paper, we described the collaborative framework named DirectSpace for supporting group workspace over Wi-Fi Direct. We adopt the Wi-Fi Direct feature as a peer-to-peer model to overcome the structural problem of client/server model. In

addition, we designed rich workspaces to effectively provide collaborative service environment in a peer-to-peer network, proposing useful services and protocols for collaborative workspaces as well.

Based on the services and protocols, we developed DirectSpace. A group workspace provides access control mechanism for resource sharing, and the workspace is managed physically as a folder on the external storage such as SD Card. Sharing and managing of resources is easily supported through the feature like as the file explorer and permission policy based on the access control mechanism. Even when mobile networks become unstable in disaster situations such as earthquakes, users can communicate through the developed useful collaborative environment.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012-0004747).

References

1. Introduction Android4.0, http://www.android.com/about/ice-cream-sandwich/
2. WiFi Direct Write Papers, http://www.wi-fi.org/sites/default/files/downloads-registered/wp_Wi-Fi_Direct_20101022_Consumer.pdf
3. WiFi Alliance, Wi-Fi Peer-to-Peer (P2P) Technical 7 Specification, http://www.wi-fi.org/Wi-Fi_Direct.php
4. Lee, H.-C., Park, J.-E., Lee, M.-J.: C3ware: A Middleware Supporting Collaborative Services over Cloud Storage. The Computer Journal (2013), WiFi Direct, http://developer.android.com/about/versions/android-4.0.html
5. Warebox mobile office, https://play.google.com/store/apps/details?id=net.warebox
6. SuperBeam WiFi Direct Share, https://play.google.com/store/apps/details?id=com.majedev.superbeam

Specification of Communication Based Train Control System Using AADL

Lichen Zhang[1] and Bingqing Xu[2,*]

[1] Faculty of Software Engineering Institute, East China Normal University, Shanghai 200062, Shanghai, China
zhanglichen1962@163.com
[2] Software Engineering Institure, East China Normal University, 200062 Shanghai, China
xbqjoya@gmail.com

Abstract. The development of railway cyber physical systems is a challenging process. In this paper we present our current effort to extend AADL to include new features for separation of concerns of railway cyber physical systems, we extend AADL in spatial aspect, dynamic continuous aspect, physical world modeling aspect. Finally, we illustrate the proposed method via an example of specification of communication based train control system.

Keywords: AADL Railroad system, Train control system, component, OSATE.

1 Introduction

The problems that must be addressed in operating a railway are numerous in quantity, complex in nature, and highly inter-related [1-3].Because of the timeliness constraints, safety and availability of train systems, the design principles and implementation techniques adopted must ensure to a reasonable extent avoidance of design errors both in hardware and software. Thus a specific methodology relevant, to design should be applied for train control systems development. The dependability of the railway cyber physical system should arouse more attention [4-5].

In this paper, we propose a specification for communication based train control system(CBTC)using AADL .We use the AADL to specifying each subsystem, and make an effective integration of all subsystems together to form a complete CBTC system.

2 The Proposed Specification Method for Communication Based Train Control Systems

AADL [6-7] is an architecture description language developed to describe embedded systems is shown in Fig.3. AADL (Architecture Analysis and Design Language),

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_10,

which is a modeling language that supports text and graphics, was approved as the industrial standard AS5506 in November 2004. Component is the most important concept in AADL. The main components in AADL are divided into three parts: software components, hardware components and composite components. Software components include data, thread, thread group, process and subprogram. Hardware components include processor, memory, bus and device. Composite components include system.

In its conformity to the ADL definition, AADL provides support for various kinds of non-functional analyses along with conventional modeling [8]:

Flow Latency Analysis: Understand the amount of time consumed for information flows within a system, particularly the end-to-end time consumed from a starting point to a destination.

Resource Consumption Analysis: Allows system architects to perform resource allocation for processors, memory, and network bandwidth and analyze the requirements against the available resources.

Real-Time Schedulability Analysis: AADL models bind software elements such as threads to hardware elements like processors. Schedulability analysis helps in examining such bindings and scheduling policies.

Safety Analysis: Checks the safety criticality level of system components and highlights potential safety hazards that may occur because of communication among components with different safety levels.

Security Analysis: Like safety levels, AADL components can be assigned various security levels.

AADL defines two main extension mechanisms: property sets and sublanguages (known as annexes). It is possible to extend the AADL concepts either by introducing new properties to the modeling elements, by addition of new modeling notations, or by developing a sublanguage as annex to the AADL [9-10].

Physical World Aspect: Railway cyber physical systems are often complex and span multiple physical domains. Modelica [11-12] is a new language for hierarchical object oriented physical modeling which is developed through an international effort. The language allows the user to specify mathematical models of complex physical systems.

Dynamic Continuous Dynamics Aspect: Railway cyber physical systems are mixtures of continuous dynamic and discrete events. These continuous and discrete dynamics not only coexist, but interact and changes occur both in response to discrete, instantaneous, events and in response to dynamics as described differential or difference equations in time.

Spatial Aspect: The analysis and understanding of railway cyber physical systems spatial behavior – such as guiding, approaching, departing, or coordinating movements is very important.

3 Case Study: Specification of Communication Based Train Control Systems

It is known that railway system can be effectively split into four subsystems: automatic train supervision subsystem (ATS), zone control subsystem, vehicle on-board subsystem and data communication subsystem. The CBTC system's file structure is expreseed by AADL ss shown in Fig.1.

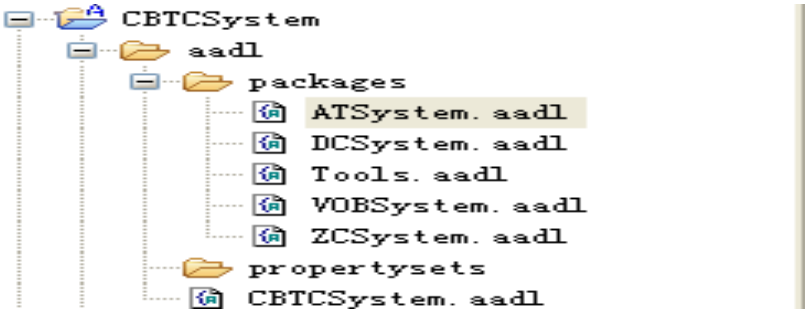


Fig. 1. CBTC system's file structure

The AADL text description of code of CBTC system is as follows:

```
System railway cyber physical LSystem
end railway cyber physical System;
system implementation railway cyber physical System.Impl
  subcomponents
    ATSys: system ATSystem::ATSys.Impl;
    ZCSys: system ZCSystem::ZCSys.Impl;
    VOBSys: system VOBSystem::VOBSys.Impl;
    DCSys: system DCSystem::DCSys.Impl;
  connections
    conn1: bus access ATSys.toDCS -> DCSys.fromATS;
    conn2: bus access ZCSys.toDCS -> DCSys.fromZC;
    conn3: bus access VOBSys.ToDCS -> DCSys.fromVOBS;
    conn4: bus access DCSys.toATS -> ATSys.fromDCS;
    conn5: bus access DCSys.toZC -> ZCSys.fromDCS;
conn6: bus access DCSys.toVOBS -> VOBSys.FromDCS;
…
End railway cyber physical System.Impl;
```

The components in the ATS subsystem are shown in Fig. 2.

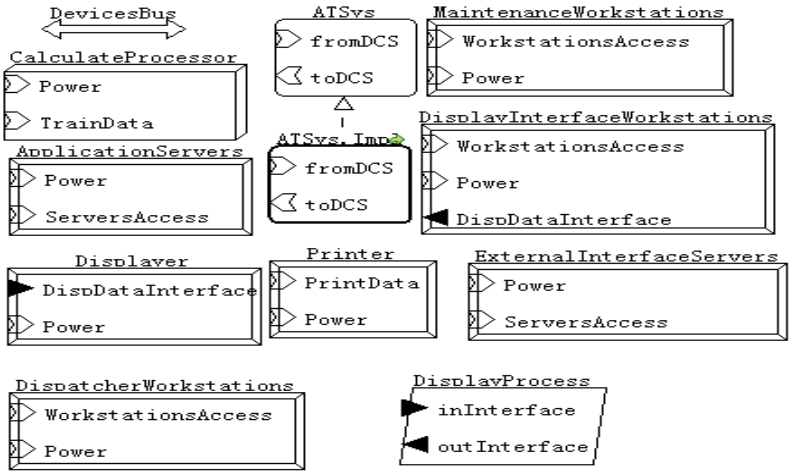


Fig. 2. The components in the ATS subsystem

We transform Modelica models to AADL property sets as Fig. 3 shows.

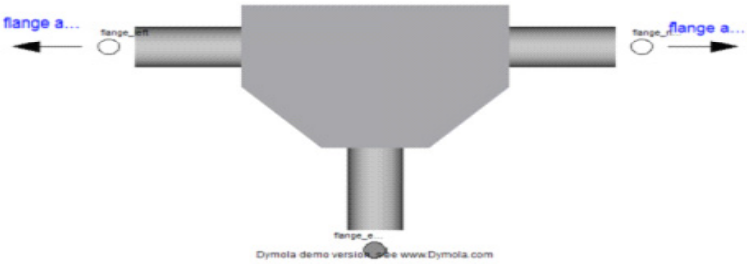


Fig. 3. Train wheel modelica model

We transform the Modelica model into AADL property sets as follows:

```
process implementation process.gearbox
properties
    physical_world_transformation:: comm_error_status => true;
end process.gearbox;

property set physical_world_transformation is

comm_error_status: aadlboolean applies to (process, device);
Left_phi: type range of aadlreal 0.0 mph..200.0 mph units (mph);
Right_phi: type range of aadlreal 0.0 mph..200.0 mph units (mph);
Left_tau: type range of aadlreal 0.0 mph..100.0 mph units (mph);
Right_tau: type range of aadlreal 0.0 mph..100.0 mph units (mph);
ratio: constants aadlinteger => 3;

end physical_world_transformation;

annex physical_envi Model{**
engine_phi=(value(physical_world_transformation::Left_phi)+
            value(physical_world_transformation::Right_phi))
            *value(physical_world_transformation::ratio)/2;
…
…
**};
```

The end to end flow analysis result of ATS is shown in Fig. 4.

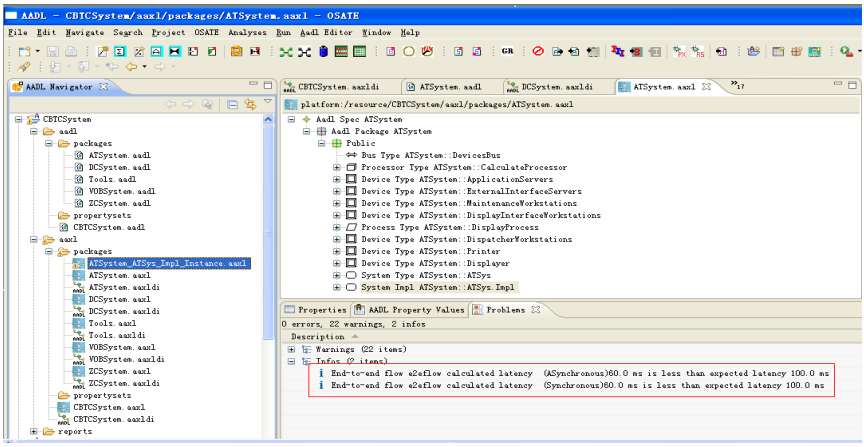


Fig. 4. The end to end flow analysis result of ATS

4 Conclusion

In this paper we presented our current effort to extend AADL to include new features for separation of concerns of railway cyber physical systems, we extended AADL in spatial aspect, dynamic continous aspect, physical world modeling aspect, formal specification aspect. Finally, we illustrated the proposed method via an example of specification of communication based train control system.

The future works will focus on the intergration AADL with formal methods.

Acknowledgment. This work is supported by Shanghai Knowledge Service Platform Project (No.ZF1213), national high technology research and development program of China (No.2011AA010101), national basic research program of China (No. 2011 CB302904), the national science foundation of China under grant (No. 61173046, No.61021004, No.61061130541, No.91118008), doctoral program foundation of institutions of higher education of China (No.20120076130003), national science foundation of Guangdong province under grant (No.S2011010004905).

References

1. IEC62278:2002, Railway applications: Specification and demonstration of reliability, availability, maintainability and safety (RAMS)
2. IEC62279:2002, Railway applications: Communications, signaling and processing systems–Software for railway control and protection systems
3. EC62280:2002, Railway applications: Communication, signaling and processing systems–Safety related electronic systems for signaling

4. Laprie, J.C. (ed.): Dependability: Basic Concepts and Terminology. Springer, Secaucus (1992)
5. Avizienis, A., Laprie, J.-C., Randell, B.: Dependability of computer systems: Fundamental concepts, terminology, and examples (2000)
6. Hudak, J.J., Feiler, P.H.: Developing aadl models for control systems: A practitioner's guide (2007)
7. Feiler, P.H., Gluch, D.P.: Model-Based Engineering with AADL: An Introduction to the SAE Architecture Analysis & Design Language. Addison-Wesley Professional (2012)
8. Muhammad, N., Vandewoude, Y., Berbers, Y., van Loo, S.: Modelling embedded systems with AADL: A practical study. In: New Advanced Technologies, pp. 247–263. Intech (2010)
9. De Niz, D., Feiler, P.H.: Aspects in the industry standard AADL. In: Proceedings of the 10th International Workshop on Aspect-oriented Modeling, pp. 15–20. ACM press (2007)
10. Michotte, L., Vergnaud, T., Feiler, P.H., France, R.B.: Aspect oriented modeling of component architectures using AADL. In: New Technologies, Mobility and Security, NTMS 2008, pp. 1–6. IEEE (2008)
11. Modelica Association, Modelica - a unified object-oriented language for physical systems modeling. Language specification (2002)
12. Modelica Association, Modelica: a unified object-oriented language for physical systems modeling: Language. Specification Version 3.0 (2007)

An Agent Modeling for Overcoming the Heterogeneity in the IoT with Design Patterns

Euihyun Jung[1], Ilkwon Cho[2], and Sun Moo Kang[2]

[1] Dept. of Computer Science, Anyang University, 602-14, Jungang-ro, Buleun-myeon, Ganghwa-gun, Incheon, 417-833, Korea
[2] Division of Digital Infrastructure, National Information Society Agency, NIA Bldg, Cheonggyecheonno 14, Jung-gu, Seoul, 100-775, Korea
jung@anyang.ac.kr, {ikcho,etxkang}@nia.or.kr

Abstract. The Internet of Things (IoT) has been considered as a core infrastructure that provides the connectivity to anyone, anywhere, anytime and especially anything. Due to this advantage, the IoT is expected to change the whole society and to enrich people's everyday life, but there are a lot of technical issues in realizing the IoT. Among them, the heterogeneity is an urgent and essential issue that cannot be easily resolved. In this paper, we described an agent modeling that can hide the heterogeneity of devices using the Strategy, Dependency Injection, and Reflection design patterns. The designed agent was implemented as the agent system named iSilo and various devices were developed and bound to the agents in the iSilo. Several experiments were conducted in Korea and Japan and these evaluations showed the proposed modeling could be a novel solution to overcome the heterogeneity in the IoT.

Keywords: IoT, Agent, Design Patterns.

1 Introduction

A lot of research groups forecast that the Internet of Things (IoT) would be an essential infrastructure which changes the whole industry and society by the roots [1][2]. The IoT is defined as a ubiquitous computing environment where various physical things such as RFID embedded objects or sensors interact with each other and harmonize together for smart IT services. However, there has been no common solution to realize it yet, because the IoT is still in the beginning stage [3].

Although there are a lot of issues to be resolved in the IoT realization, the most essential one is the heterogeneity because the IoT needs a close collaboration of various kinds of things having different hardware, identifiers, network connectivity, and interworking protocols. That is, wireless sensors, RFID tagged products, home appliances, and all other mixed type devices should identify and communicate with others for the IT services promised by the IoT. Unfortunately, in the existing technologies, there is no way to overcome the heterogeneity successfully.

In order to resolve the heterogeneity issue in the IoT, the researchers of Cyber-Physical System (CPS) have paid attention to an agent technology and they believed

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_11,

that the agent technology would be a solution to hide nicely the heterogeneity [4]. In the proposed scenario, devices delegate their authority to the corresponding agents and actual collaboration is done in the agent level. Since the agents identify and communicate with other agents in the agent space, the differences of physical devices can be easily hidden. This seems a good idea to use the agent technology for overcoming the heterogeneity, but the detailed structure or the reference model was not represented yet.

Basically, to hide the heterogeneity, an agent has to implement every different function of various devices in it. It seems an intuitive approach, but it is impractical for the IoT because every function of devices cannot be anticipated and implemented in the development phase. Additionally, even if this approach is adopted, the agent must be modified in the source code level and be redeployed again into the system whenever a new device is appeared or the functions of a device are changed. Therefore new approach with which an agent dynamically changes itself depending on the types of devices should be considered.

In order to fulfil the requirement, we designed an agent with several design patterns; Strategy, Dependency Injection (DI), and Reflection. The designed agent separates its core and the device-dependent functions. At start time, the agent attaches the needed functions to itself in order to change its behavior for the correspondent device. We also developed an agent service platform named iSilo to evaluate the proposed agent design. In the iSilo, various devices with different sensing functions and network connectivity delegate their authority to the corresponding agents. We conducted several experiments of the iSilo after deploying the devices made with Arduino [5] in both Korea and Japan. The experiments showed that the same agent could handle totally different devices without modifying itself and various devices could collaborate through the cooperation of their agents. This evaluation indicates that the proposed agent model could be a novel method to overcome the heterogeneity in the IoT.

The remainder of this paper is organized as follows: The adoption of design patterns to the structure of an agent is stated minutely in Section 2. In Section 3, the implementation of the agent is described and it is evaluated to show the effectiveness. Finally, Section 4 gives the conclusion.

2 Agent Modeling

2.1 Portable Service Abstraction with the Strategy Design Pattern

In order to support the every function of the various devices, the separation of the functions from the agent core and the dynamic binding is required. However, this separation leads to a new problem that an agent core has to know all method signatures of the device-dependent functions. That is, the agent core has no choice but to contain the name of all classes and methods as shown in Fig.1.(a), so a simple separation cannot be a solution in considering the vast diversity of IoT things.

This problem has been referred as the dependency in the Object Oriented Design (OOD) and the Portable Service Abstraction (PSA) policy has been proposed to reduce the dependency between objects. The PSA recommends architects to use an interface rather than a concrete instance for achieving low coherence between objects.

As shown in Fig.1.(a), the agent core has to know object references and every signature to call the needed functions. In this structure, when a new device is added, the agent core should be modified in the source code level and be redeployed. In our design shown in Fig.1.(b), all methods are abstracted as the Behavior interface. Then, the agent core can call various objects using the same Behavior interface if the classes implement the Behavior interface. Unlike the conventional system, when a new device is added, it will be easily supported without modifying the agent core. The resulted structure is the form of the Strategy design pattern. With the pattern, the agent core can act as a proxy of various devices by equipping a device-dependent class which implements the Behavior interface.

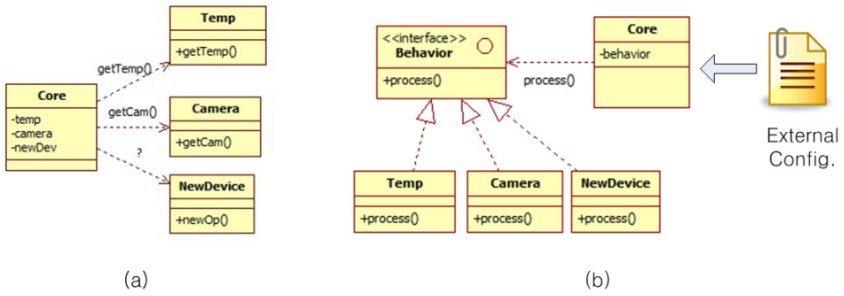


Fig. 1. The UML diagram of the conventional structure (a) and the proposed structure (b)

2.2 The Dependency Injection

The Strategy pattern provides a way to equip the target class having device-dependent functions without modifying the agent core, but the agent core still has to hold the target class name in the source code because the instantiation of the target class is unavoidable. We resolved this problem using the Dependency Injection (DI) of the Spring framework. The DI is a design pattern that allows to remove hard-coded dependencies in the source code and to inject them in the external configuration. The DI ensures that the agent core does not know where device-dependent functions come from and it allows the selection among multiple implementations of the given interface at runtime. Using the DI and the Strategy patterns, the agent core can choose the proper equipment classes from the external configuration as shown in Fig.1.(b) without knowing their actual class names.

2.3 The Reflection

By adopting the DI and the Strategy patterns, the agent core can assemble the needed functions without worrying about the dependency problem. However, the proposed design still has a limit that a single method signature is commonly used for the different type of functions. An agent of the IoT has to handle the different category of functions which should be optimized for various types of devices. For example, the network function and the sense function may require the different method signature

for their own purpose, but the proposed design forces the target classes to implement the identical interface and it unavoidably results in limiting the extensibility.

We studied whether it was possible to call the different method signature while maintaining the same interface and we finally concluded that the Reflection could be a solution for it. The Reflection makes it possible to inspect classes, interfaces, fields and methods at runtime, without knowing the names of the classes, methods etc. at compile time. It is also possible to invoke methods by delivering its signature string. Using the Reflection, the same interface can deliver the different method signature to its functions. In order to use the Reflection in the Spring framework, AspectJ [6] and Dynamic Proxy patterns [7] are used. After the implementation, network, power, and sense functions are called with their own method signatures via the identical interface.

3 Implementation and Evaluation

In order to evaluate the proposed design, an agent service platform named iSilo was implemented on the Spring framework. Also, several devices having various sensing functions and different network connectivity were developed with Arduino to emulate real world things. The developed devices were 11 different kinds of IoT devices; a temperature sensor, a humidity sensor, a human presence detector, a vibration sensor, a radioactive sensor, a camera sensor, a rainfall meter, a wind vane, a wind gauge, a speaker and an illuminometer. These devices were located at Anyang and Pusan university of Korea and at Waseda university of Japan and they were connected in three kinds of network; Ethernet, Wi-Fi, and ZigBee. As it was designed, the agent in the iSilo acted as a proxy of each device and cooperated with other agents.

3.1 Agent's Dynamic Equipment of Device-Dependent Functions

In modeling the agent, we focused on the agent's dynamic support of the device-dependent functions without modifying the source code of the agent core. To evaluate this design purpose, we conducted an experiment in which the two same kinds of agents act as a proxy of two different kinds of sensors; a camera sensor connected through the Ethernet and a temperature sensor connected through the ZigBee.

The camera sensor reports binary image data to the corresponding agent, which stores the data into mass storage and renders the stored images to the management console. On the other hand, the agent of the temperature sensor stores the reported temperature data into a database and it renders the data as a bar chart. As shown in Fig. 2, the each agent needs not to modify the agent core at all, but just to indicate proper device-dependent functions in the external configuration. This experiment shows any kind of device can be easily supported by the proposed agent without the code modification if the corresponding device-dependent function is properly indicated in the external configuration.

After the experiment, we changed a report function from "CelsiusBehavior" class to "FahrenheitBehavior" class in the configuration of the agent for the temperature sensor. This assumed the situation where a device changed its behavior after finishing

the deployment of IoT services. In this sub experiment, the reported value is well processed without affecting other parts of the system and even the corresponding agent itself.

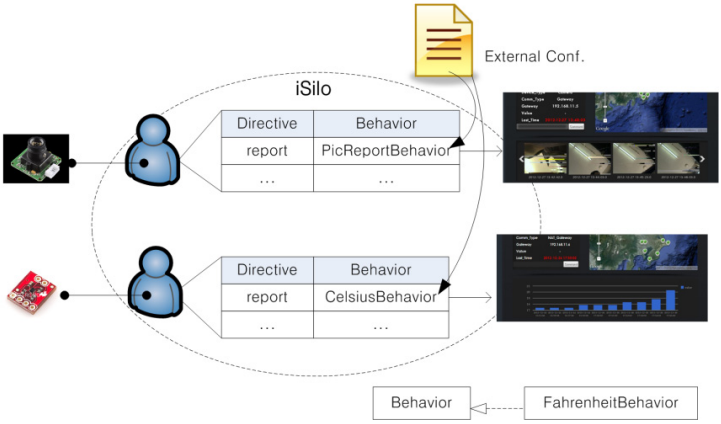


Fig. 2. The structure of Agent's Dynamic Equipment for heterogeneous devices

3.2 Collaboration between Heterogeneous Devices

Another design purpose of the proposed structure is the collaboration of heterogeneous devices. In order to evaluate the design purpose, a radioactive sensor connected through WiFi in Japan and a speaker actuator connected through ZigBee in Korea were selected. We wanted that the radioactive sensor would notify an event to the speaker when the sensor detected the abnormal level of radioactivity. Needless to say, this notification between heterogeneous devices was impossible with current technology.

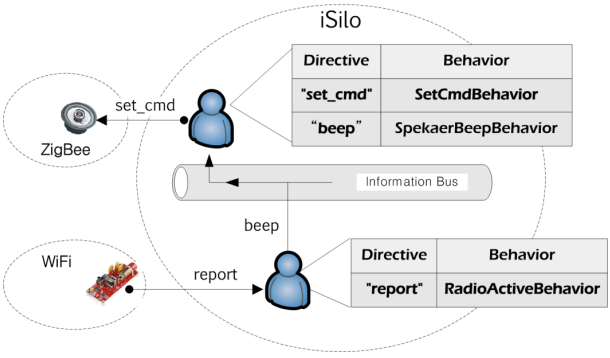


Fig. 3. The structure of device collaboration by agent communication

However, as shown in Fig. 3, this notification was very successful in the experiment. When an experimenter placed radioactive material near the radioactive sensor, the sensor reported the radioactive level to the radioactive delegating agent through

WiFi. Then, the corresponding agent sent the event to the speaker delegating agent through the information bus in the iSilo. After receiving the event, the speaker delegating agent had the speaker make a beep sound. This showed that various collaborations of heterogeneous devices are possible through the agent communication in the proposed design.

4 Conclusion

The IoT has attracted a lot of research groups and it has been considered as a core infrastructure for the future ICT. However, there are many obstacles to realize the IoT. Among them, the heterogeneity is the most essential issue and some researchers believe that an agent technology would be a solution. The idea is that the devices of the IoT delegate their authority to agents and these devices collaborate with each other via their agents. However, a detail structure or a reference model for the agent has not been presented yet.

In this paper, we suggest an agent modeling which resolves the heterogeneity issue by using the design patterns. Generally, putting all functions of devices into a single agent is impractical and error prone. Therefore, we adopted the Strategy, the DI, and the Reflection patterns to separate the agent core and the device-dependent functions. These patterns contributed not only to reduce the dependency and but also to enable flexible function extensions. The prototype implementation and the experiments showed that the proposed agent could handle heterogeneous devices without modifying itself and it could provide the collaboration of heterogeneous devices.

References

1. Bandyopadhyay, D., Sen, J.: Internet of Things: Applications and Challenges in Technology and Standardization. Wireless Personal Communications 58(1), 49–69 (2011)
2. Atzori, L., Iera, A., Morabito, G.: Internet of Things: A Survey. Computer Networks 54(15), 2787–2805 (2010)
3. Sundmaeker, H., Guillemin, P., Friess, P., Woelffle, S.: Vision and Challenges for Realising the Internet of Things. Cluster of European Research Projects on the Internet of Things, European Commision (2010)
4. Lin, J., Sedigh, S., Miller, A.: A Semantic Agent Framework for Cyber-Physical Systems. In: Elçi, A., Koné, M.T., Orgun, M.A. (eds.) Semantic Agent Systems. SCI, vol. 344, pp. 189–213. Springer, Heidelberg (2011)
5. Wilcher, D.: Learn Electronics with Arduino. Apress (2012)
6. Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., Griswold, W.G.: An Overview of AspectJ. In: Lindskov Knudsen, J. (ed.) ECOOP 2001. LNCS, vol. 2072, pp. 327–354. Springer, Heidelberg (2001)
7. Fox, A., Gribble, S., Chawathe, Y., Brewer, E.: Adapting to network and client variation using active proxies: Lessons and perspectives. IEEE Personal Communications 5(4), 10–19 (1998)

BK-means Algorithm with Minimal Performance Degradation Caused by Improper Initial Centroid

Hoon Jo[1,*] and Soon-cheol Park[2]

[1] Division of Electronics and Information Engineering Department,
Chonbuk National University, Jeonju, Korea
ei201250226@jbnu.ac.kr
[2] Division of Electronics and Information Engineering Department,
Chonbuk National University, Jeonju, Korea, and IT Convergence Research Center
scpark@chonbuk.ac.kr

Abstract. K-means algorithm has the performance degradation problem due to improper initial centroids. In order to solve the problem, we suggest BK-means (Balanced K-means) algorithm to cluster documents. This algorithm uses the value, α, to adjust each cluster weight which is first defined in this paper. We compared the algorithm to the general K-means algorithms on Reutor-21578. The experimental results show about 11% higher performance than that of the general K-means algorithm with the balanced F Measure (BFM).

Keywords: Clustering, Information Retrieval, K-means, BK-means, Outlier.

1 Introduction

Internet produces tremendous data and information every day. It becomes to need to cluster the data according to their characteristics. K-means algorithm, one of the clustering algorithms, has been widely used in various fields because of the characteristics of easy implementation and fast clustering speed.

However, K-means has the weakness that the performance is heavily depended on the initial centroids [1]. In order to solve the problem, we proposed BK-means (Balanced K-means) algorithm to cluster documents. In this algorithm the value, α, is applied to adjusting each cluster weight. The experimental results show that the algorithm overcomes the initial centroid problem of the K-means algorithm.

This paper is organized as follows. The next section introduces the BK-means algorithm. Section 3 presents the experimental setting, results and analysis. Section 4 concludes and discusses future works.

2 BK-means

BK-means algorithm which is based on K-means algorithm uses appliance of weight to each cluster to improve under-performance caused by improper initial centroid.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_12,

2.1 K-means

K-means algorithm is an unsupervised clustering algorithm. It decides the object where it belongs in the cluster. During the process, the K-means algorithm measures the similarity between the object and cluster.

Pseudo-code of overall follow for K-means algorithm is as follows:

```
K-means Algorithm
```
1. *Arbitrarily choose an initial k centers* $C = \{c1, c2, \ldots, ck\}$
2. *For each* $i \in \{1, \ldots, k\}$, *set the cluster* C_i *to be the set of points in* χ *that are closer to* c_i *than they are to* c_j *for all* $j \neq i$
3. *For each* $i \in \{1, \ldots, k\}$, *set the* c_i *to be the center of mass of all points in* C_i
4. *Repeat steps* 2 *and* 3 *until* C *no longer changes.*

K-means algorithm decides the initial centroid as the step1. After the initial centroid is decided, each object decides the cluster through using the similarity between each cluster and object as shown in step2. After the clusters about all objects are decided, each centroid is recalculated for each cluster as step3.

In case of termination condition is satisfied, this algorithm is terminated through step4. If termination condition is not satisfied, it repeated step 2 and 3.

2.2 BK-means

BK-Means algorithm is calculated using each virtual centroid given with different weights to each cluster. The virtual centroids are centroids applied with each weight. BK-means algorithm is algorithm for improving degradation of the performance when improper initial centroid is selected.

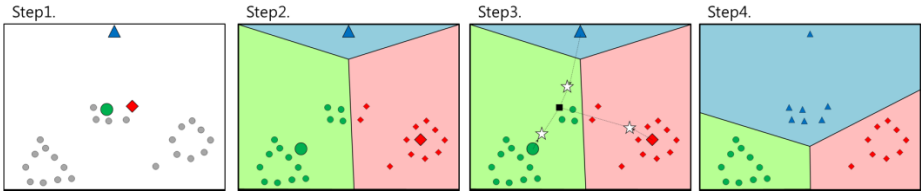


Fig. 1. Overall clustering process of BK-means

Overall clustering process of BK-means is as shown is Fig.1. In Fig.1, small (▲●◆) means the objects which are included in the each cluster. Big (▲●◆) means the centroids of each cluster. In Step 1, (●) means the initial object which is not clustered. In step 3, (☆) means virtual centroid. And (■) is the object which will be clustered. Step 1 shows the process which decides initial centroid. In step1, (▲) has a propensity of outlier.

If initial centroid is determined, this algorithm determines each cluster through calculating the similarity between each object and cluster. And then, centroid is moved to the center of the cluster. This centroids are replaced to virtual centroids applied with different weight of each cluster as Step3.

Equation of Weight is as follows:

$$\text{Weight}_{\text{i}} = (\log\frac{N-nC_i}{N}) \times \alpha \tag{1}$$

In equation.1, $\boldsymbol{N}$ is number of all vectors. $\boldsymbol{nC_i}$ is number of vector of $\boldsymbol{Class_i}$. $\boldsymbol{\alpha}$ is variable between 0 and 1. α is doing very important role as a factor because it determines how much influence to the weight. The virtual centroid means centroid by the result of calculating between each cluster and weight.

Similarity equation for virtual centroid is as follows:

$$ModCosSim(C_i, x) = \frac{\vec{V}(C_i)\cdot\vec{V}(x)}{|\vec{V}(C_i)||\vec{V}(x)|} + Weight_i \tag{2}$$

In equation.2, $\frac{\vec{V}(C_i)\cdot\vec{V}(x)}{|\vec{V}(C_i)||\vec{V}(x)|}$ means common cosine similarity.

3 Experiment

Table 1 is a dataset for experiment. In Table1, classes mean each cluster. The each element of the classes means the labels which represent their cluster. And the number in parentheses is the number of document.

Table 1. Dataset for experiment

	Class1	Class2	Class3	Class4
Set1 (20NG)	comp.graphics (200)	rec.motorcycles (200)	sci.crypt (200)	-
Set2 (Reuter)	crude (200)	earn (200)	grain (200)	-
Set3 (20NG)	rec.sport.baseball (200)	sci.electronics (200)	soc.religion.christian (200)	talk.politics.guns (200)
Set4 (Reuter)	interest (200)	money-fx (200)	trade (200)	wheat (200)
Set5 (20NG)	sci.space (300)	comp.sys.ibm.pc.hardware (200)	rec.sport.hockey (100)	sci.med (50)
Set6 (Reuter)	acq (300)	crude (200)	money-supply (100)	ship (50)

Reuter which is most commonly used includes 21,578 news articles as Reuter-21578 data collection. 20NG consists of each 1000 data in 20 UseNet group as 20 Newsgroup data collection [2].

To evaluate the effectiveness of the proposed method, we used balanced F Measure (BFM). BFM means giving equal weight to Precision and Recall.

Equation of BFM is as follows:

$$F_{\beta=1} = \frac{2*Precision*Recall}{Precision+Recall} \tag{3}$$

In our experiments, we used Document Frequency (DF) method for improving the performance. When feature has high-frequency, DF method which is one of the kinds of Feature Selection (FS) determines that feature is the most important [3].

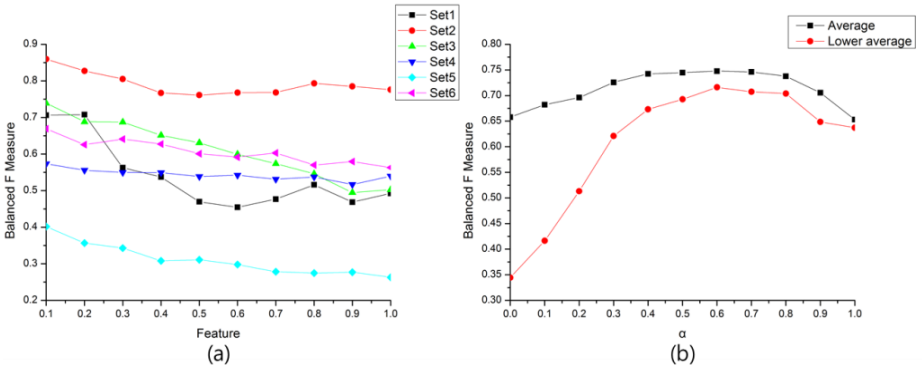


Fig. 2. Balanced F Measure according to the feature and value of α

Fig.2 (a) which used Table1's dataset by FS shows variation of BFM. Range of FS is 0.1 to 1.0. Meaning of 1 is using all vector of document. As a result, K-means algorithm shows the best performance when using the FS of the highest 0.1(10%). Fig.2 (b) shows the change of efficiency measured as value of α of weight when using the feature of the highest 10%. When value of α is 0, it generally means K-means. In Fig.2 (b), 'Average' is an average variation of Set1~ 6 as a change of value of α. BFM of BK-means is 11% higher than K-means when value of α is 0.6 which is the optimal value of α. 'Lower Average' measures the average variation through using BK-means about dataset which shows the lowest 10 % performance with K-means. When value of α is 0, BFM shows average performance of 34%. When value of α is optimal value, performance increases from 34% to 74%.

Table 2. Optimal α for each dataset

	Set1	**Set2**	**Set3**	**Set4**	**Set5**	**Set6**
Optimal α	0.4	0.6	1	1	0.9	0.4

Table 2 is optimal value of α for each dataset. Optimal value of α is value of α which shows the highest BFM when using the highest 10% FS about each dataset.

Fig.3 is variation of performance when using optimal value of α with Set1~6 in Table 1. In Fig.3, (a) is the results of K-means and BK-means about dataset using feature of the highest 10% by FS. (b) is the results about using dataset showing the performance of the lowest 10% by K-means.

From Fig.3 (a) and (b), we can see that experiment performance of BK-means which is taken on 10% lowest of K-means increased close to general average performance.

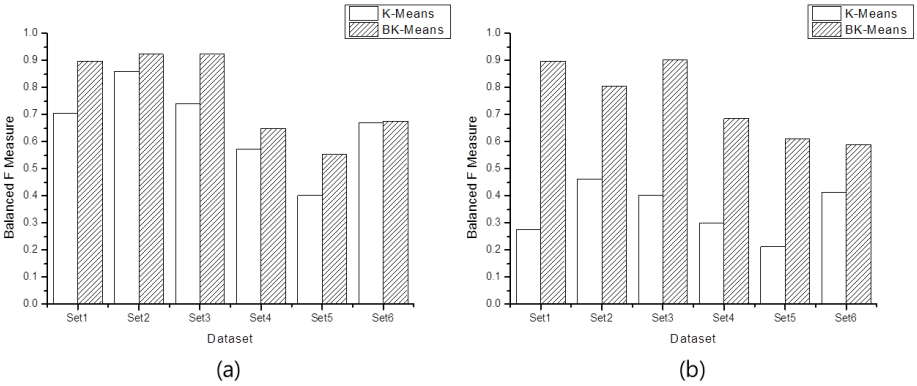


Fig. 3. Variation of performance according to the optimal value of α

4 Conclusion

In this paper, we propose the BK-means algorithm which show the minimal performance degradation caused by improper initial centroid, not as in the general K-means algorithm. The BK-means algorithm uses the virtual centroids to avoid that an outlier become one of the centroids. This solves the improper initial centroid problem and makes to improve the clustering performance. This algorithm shows in general about 11% higher performance than the K-means algorithm with the balanced F-measure. Especially, it shows about 40% higher performance than the K-means algorithm does for the data which shows the bad clustering performances with the K-means algorithm. It means that the lowest-performance is improved nearly average performance when BK-means is used.

In our experiment, the BK-means algorithm has the highest performances when the value, α, adjusting the centroid weights is around 0.6. In the future, the method to find the optimal value of α for the document clustering algorithm will be researched precisely for the higher performance.

Acknowledgements. Following are results of a study on the "Leaders in INdustry-university Cooperation(LINC)" Project, supported by the Ministry of Education, Science & Technology(MEST) and the National Research Foundation of Korea (NRF - No.1301000393).

References

1. Arthur, D., Vassilvitskii, S.: K-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035. Society for Industrial and Applied Mathematics Philadelphia (2007)
2. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
3. Liu, T., Liu, S., Chen, Z., Ma, W.Y.: An evaluation on feature selection for text clustering. In: Proceedings of the Twentieth International Conference, pp. 488–495. AAAI Press, Washington, DC (2003)

Detect Spatial and Temporal Gait Parameters by Dual Accelerometers

Wann-Yun Shieh, An-Peng Liu, and Tyng-Tyng Guu

Department of Computer Science and Information Engineering,
Healthy Aging Research Center,
Chang Gung University, Taiwan
wyshieh@mail.cgu.edu.tw

Abstract. The process of walking or running is called the "gait". In clinical research, gait detection can be used to investigate the features of normal or abnormal gait for demonstrating a change from treatment or from disease progression. In the past, many optical-based gait detection approaches have been proposed. In these approaches, we have to paste many reflective markers on the subject's limbs and use multiple cameras from different directions to take the images of walking. They can provide high accuracy measurements for gait detection, but they also need very expensive optical equipment. Also, the experiments are restricted to the laboratory environment, which means that the collection of gait data will be limited in a short distance or a short time interval. In this paper we will propose a portable design, which uses dual accelerometers pasted on a subject's left and right waist to do the gait detection at any time, any place. Particularly, we will apply the wireless communication to develop a gateway, as well as its App on the smart phone, to collect sensing data. The data collected from the sensors can be uploaded to the remote cloud for many telemedicine applications.

Keywords: Gait detection, accelerometers, telemedicine applications.

1 Introduction

Walking is one of the most important skills in daily life. To evaluate this skill, most studies will collect the gait parameters as the indicators [2]. If a person has any problems in gait, the imbalance of standing or walking will be the first symptom to be presented. Moreover, such imbalance may cause harmful incidents, like falls. There are three major kinds of physiological factors that will cause gait imbalance. The first factor is the disease or the degradation of the nerve system, e.g., the Parkinson's disease, which will make nerves conduction slower. The second factor is the degradation of the bones, joints, or muscles. The pain or deformation of lower limbs will also affect a person's posture and then leads to the diminution in gait stability and balance. The third factor is that the patients may have lower limbs injury or hurt.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_13,

Through gait detection, we can detect the degree or classify the status of the standing imbalance or walking instability [3]. In clinical research, gait detection can also be used to investigate the features of normal or abnormal gait for demonstrating a change from treatment or from disease progression. By analyzing the gait of elderly, we can observe whether they have abnormality or degradation in walking ability. For the patients with leg or foot injury, the gait analysis can be used to evaluate the effectiveness of rehabilitation or balance assessment after a surgery operation.

To detect the gait, we can break the repetitive body movements into gait cycles, Each gait cycle can be characterized by several gait events. Take Fig. 1 for example. Fig. 1 shows a process from the posture of the right foot touching the ground to the same posture of the right foot touching the ground again. If the posture of any one foot touching the ground is defined as the "heel-strike", and the posture of any one foot leaving the ground is defined as the "toe-off", then a gait cycle can be broken into five gait events: $R_{heel\text{-}strike}$, $L_{toe\text{-}off}$, $L_{heel\text{-}strike}$, $R_{toe\text{-}off}$, $R_{heel\text{-}strike}$, as shown in Fig. 1, where "*R*" denotes the right foot and "*L*" denotes the left foot.

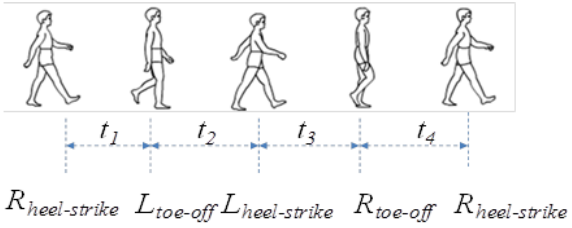


Fig. 1. The process of a gait cycle

By the definition of a gait cycle, several well-known gait parameters can be quantified as follows:

(1) Gait Cycle Time (*GCT*): The period of the time between the heel-strike of one foot to the next heel-strike of the same foot, i.e., $GCT=t_1+ t_2+ t_3+ t_4$, where t_i is the time interval between two sequential gait events in Fig. 1.
(2) Step number (N_{step}): The number of steps in a fixed distance. This parameter can be counted by each leg separately, e.g., we can count the number of left heel-strikes as the step number of the left foot.

In the past, many gait detection techniques have been proposed. Most of them are optical-based approaches, in which we have to paste many reflective markers on the subject's limbs and use multiple optical cameras from different directions to take the pictures of standing or walking. This kind of approaches can provide very accurate measurement for gait analysis, but they usually need expensive optical equipment. Also, the experiments are restricted to the laboratory environment.

An alternative approach is to use the accelerometers to perform the gait detection [1-2]. Current technology has made these sensors very small and light-weight such that they can be embedded into wearable clothes or shoes to detect motions. Moreover, most accelerometers consume much less power, which means that this kind of devices can be used at home for long-term data collection. The data collected

can be analyzed by the doctors via a remote platform. It has been shown that much gait information a subject performed at home in daily life can be exploited for significant diagnosis or prevention of motion problems.

In this paper, we apply the wearable sensors and the wireless communication technology to develop a dual-accelerometer-based smart belt, such that a subject can wear it on the waist to perform the left and right lower-limb gait detection and analysis at home. Particularly, we use this belt to design an approximation algorithm to calculate not only temporal parameters, but also spatial parameters of gait. The gait data will be collected by a smart phone or other handheld devices to perform real-time analysis and rehabilitation suggestion. On the remote side, an information cloud is constructed to allow users uploading their data, analyzing and discussing long-term records with doctors or medical experts. Experimental results show that the proposed algorithm can achieve significant precisions for gait analysis with acceptable costs.

2 Smart Belt Design

Fig. 2 shows the prototype of the proposed smart belt. We use two ADXL330 accelerometers as the gait detection sensors which are clasped on the left and right sides of the belt. The ADXL330 is a small, thin, and low-power 3-axis accelerometer with signal-conditioned voltage outputs. It can measure the static acceleration of gravity in tilt-sensing applications, as well as dynamic acceleration resulting from motions. In our prototype, the positions of the accelerometers could be adjusted according to the subject's waistline.

The signals from the sensors will be collected and processed by a data processing gateway which is clasped on the middle of the belt. Fig. 3 shows the block diagram of the circuit, which contains a microprocessor (Pic16F877), a Bluetooth module (BTM400_6B), and a power module. The microprocessor will perform the sampling and analog/digital conversion. We let the sampling rate of the signals as 100Hz, which is very sufficient for capturing the characteristics of the gait. The data will be transferred to a mobile phone or other hand-held devices via Bluetooth for real-time analysis. The power module contains a 9-volt battery which provides the whole circuit and the dual accelerometers with sufficient electronic power. By such a waist-belt design, we do not need duplicated processors or batteries with sensors to be clasped on the left and right knees or ankles. It means that we could significantly reduce the hardware costs, as well as the software complexity on signal synchronizations.

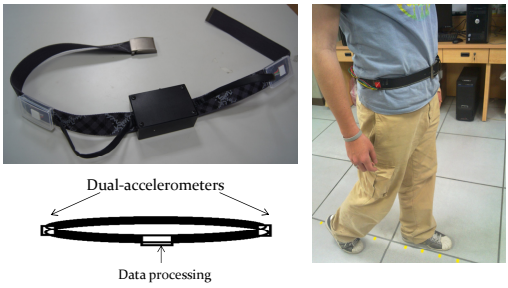


Fig. 2. The prototyping of the smart belt

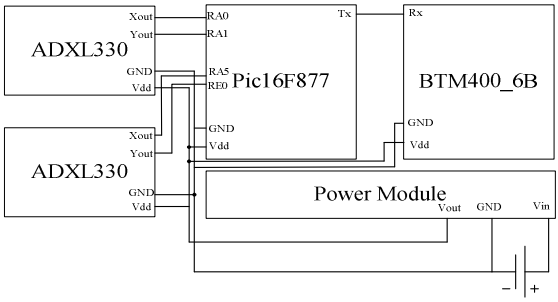


Fig. 3. The block diagram of the smart belt

Fig. 4 shows the left-side and the right-side signals of the continuous accelerations captured by the smart belt during a short period of walking (about six steps). We can find that the signals have significant pulses at each step. In addition, the peak responses will appear alternatively on each side. It shows that the signals can be used not only to identify the events of the gait, compare the differences between two lower limbs, but also compute the spatial and temporal parameters about the gait. If we assume that

k : the total step number,
t_n : the time of the n-th heel-strike (left or right), where n=1, 2, …, k, and
$X(m)$: the position of the m-th time point, where $m= t_1, t_2\ldots , t_n$,

then many other spatial-temporal parameters for gait detection can be derived, which are shown in Table 1. Note that, k can be obtained by counting the total number of peak points from each accelerometer, and t_n can be obtained by recording the time at which a peak point occurs. Assume that the position of the starting point is zero. By computing the spatial parameters from the accelerations of the sensors, the relative distance of each peak point away from the starting point, i.e., $X(m)$, can be obtained accordingly.Compared with the optical-based approaches, using the accelerometers for gait detection has the benefits from their small size, low cost, and low power, which can be applied to lightweight or wearable devices brought by the subjects without any spatial or temporal restrictions.

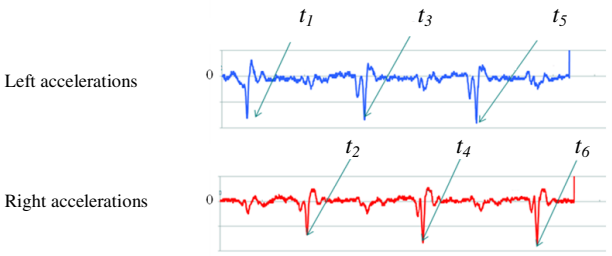


Fig. 4. The continuous accelerations from the left and right lower limbs

Table 1. Expressions of spatial and temporal parameters

Parameters	T_n	T_{step}	F_{step}	D_n	S_{step}	V_{step}
Equation	$t_n - t_{n-1}$	$\frac{\sum_{n=1}^{K}(t_n - t_{n-1})}{K}$	$\frac{1}{T_{step}}$	$X(t_n) - X(t_{n-1})$	$\frac{\sum_{n=1}^{K}[X(t_n) - X(t_{n-1})]}{K}$	$\frac{S_{step}}{T_{step}}$

Note: T_n: the period time of the n-th step; T_{step}: average period time of a step; F_{step}: average step frequency; D_n: the distance of the n-th step; S_{step}: average distance of a step; V_{step}: average speed of a step

3 Experiments

We designed two experiments to verify the effectiveness of the smart belt. In the first experiment, we let the subjects wear the belt to perform each single motion in a fixed distance. The distance we tested includes a short distance (10m~200m) and a long distance (200m~ 800m). Our goal is to verify that the belt could have precise computation in terms of the moving distance (e.g., $X(m)$) for individual motions. In the second experiment, we let the subjects randomly perform different motions in a fixed time (5mins~20mins). We also measured the total moving distance that the subject performed in each fixed time interval. Our goal is to verify that the belt could have the ability to adapt the computation along with subject's different motions. We totally select 10 volunteers whose heights range from 165cm to 175cm as the subjects. The experimental results show that for the subjects performing different motions, the smart belt could have 9.7% error rate in the fast walking motions and 7.6% error rate in the slow walking motions on average.

4 Conclusion

In this work we propose a smart belt design to detect the left and right lower-limb signals. The smart belt can be applied to many homecare services. For example, if a subject is joining a low-limb rehabilitation plan at home, this belt can be used for real-time gait detection, training guiding, and falling incident alarming etc.

References

1. Salehi, M., et al.: A sensor-based framework for detecting human gait cycles using acceleration signals. In: Proceedings of the 17th International Conference on Software, Telecommunications & Computer Networks, pp. 328–332 (2009)
2. Zijlstraet, W., et al.: Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. Gait & Posture 18(2), 1–10 (2003)
3. Ran, Y., et al.: Applications of a Simple Characterization of Human Gait in Surveillance. IEEE Transactions on Systems-Part B: Cybernetics 40(4), 1009–1020 (2010)

Implementation of Load Management Application System in Energy Management Service*

Taekyeong Kang[1], Hyungkyu Lee[2], Dong-Hwan Park[2], Hyo-Chan Bang[2], and Namje Park[1,**]

[1] Major in Elementary Computer Education, Department of Primary Education, Graduate School of Education, Jeju National University, 61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Korea
{ktg,namjepark}@jejunu.ac.kr
[2] Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, Korea
{leehk,dhpark,bangs}@etri.re.kr

Abstract. As the Smart grid is intelligent power grid, combining information Technology to the existing power grid. Electricity suppliers and consumers exchange real-time information to two-way and is a next-generation power grid to optimize energy efficiency. This paper suggests the implementation of load management application system in energy management service environment.

Keywords: Energy Management, Load Management, Smart Socket, Energy.

1 Introduction

Load leveling is method for reducing the large fluctuations that occur in electricity demand, for example by storing excess electricity during periods of low demand for use during periods of high demand. The load management system compares total power consumption and quarterly power consumption of digital meter centering on cabinet panel and sends the "interruption" command to the sockets mounted to the power source in the order of priority when the power consumption exceeds the quarterly limit in order to control power which is not urgently required or not required at all.

There are similar systems such as the power monitoring system which detects the power currently consumed by the user as well as the power consumed during a certain period to induce better power conservation, the power saving system which minimizes power consumption for unnecessary use by reserving the consumption of power by home appliances such as the management of power consumption through the out mode or ordinary mode depending on whether the user is inside or outside the

* This work was supported by the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy (MKE, Korea) [10038653, Development of Semantic Open USN Service Platform].

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_14,

building, the added service system minimizes unnecessary use of the electricity in the time zone when the demand of power is low by responding to user demand when the flexible rate system which changes rates for different time zones would be applied, and the temporary system for selling the power generated through solar power generation equipment to the power companies.

2 Overview of Load Management System

The load management system can be generally divided into 4 units: smart cabinet panel, smart socket, data receiver, and load management program as in Figure 1. The smart socket and smart cabinet panel send and receive data through Zigbee communication as in Figure 1 and the data receiver and load management program, UART (Universal Asynchronous Receiver/Transmitter) communication to monitor the information coming from smart socket and smart cabinet panel and control the device.

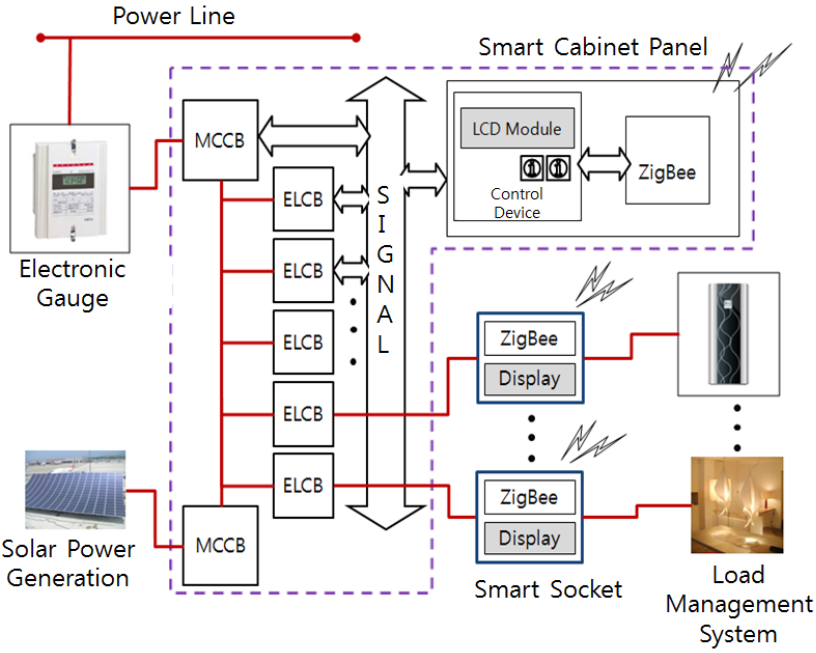


Fig. 1. System conceptual diagram of LMS

3 Implementation of Load Management Application System

Load management application program allows the user to control and monitor each device. The interface for load management application program is shown in Figure 2.

The screen for load management application program is generally divided into the device connection information area (A) and device control and monitoring area (B). In Area A, set the load management program and data receiver information and make

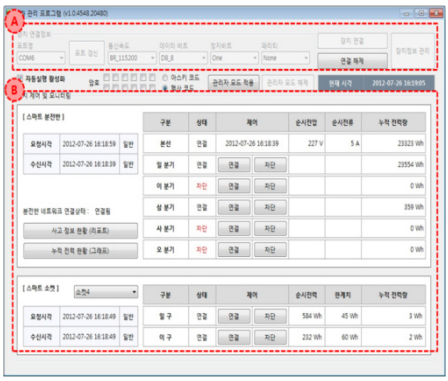


Fig. 2. Load Management Application Program Interface

connection. Area B is basically disabled. When the meta-information of smart cabinet panel and smart socket is registered through device information management for device connection, the data receiver is connected, and the device connection in Area A succeeds, Area B will be activated. Area B is divided into the smart cabinet panel on the top and smart socket on the bottom. The right side of Area B is allocated for real time data monitoring and control. The previously stored information can be monitored and analyzed through the accident information status and accumulated power status on the bottom left side of smart cabinet panel. In the smart socket area, sockets can be selected with the socket selection combo boxes on the top left for monitoring and control of the selected sockets.

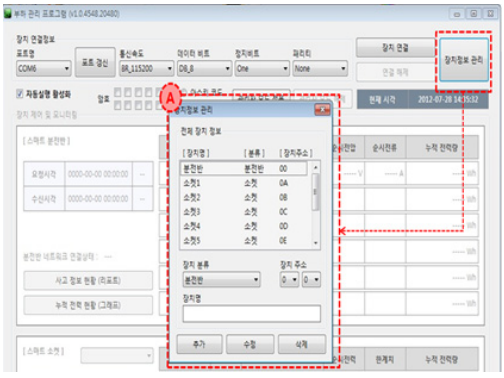


Fig. 3. Device information management display

Click the Device Information Management on the top right side of the main monitoring screen to open the management screen. Device information management is shown in Area A. The top side of Area A shows the list of meta-information of all devices and the detailed information will be displayed when the corresponding device is selected. Manage the meta-information of the device by using "Add", "Edit", and "Delete" buttons on the bottom of the screen. The load management program can operate at least one cabinet panel device information is registered.

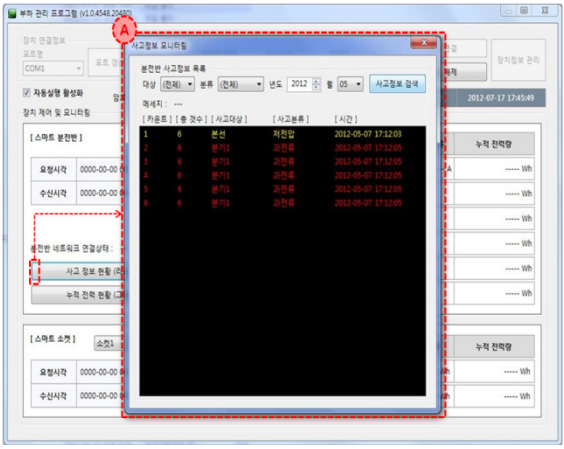


Fig. 4. Accident information monitoring display

Click "Accident Information Status" button of the smart cabinet panel in the main screen to start "Accident Information Monitoring." "Accident Information Monitoring" is shown in Area A. Search for accident information by selecting "Subject", "Category", "Year", and "Month" options on the top and the search result will appear in the form of list of text on the bottom.

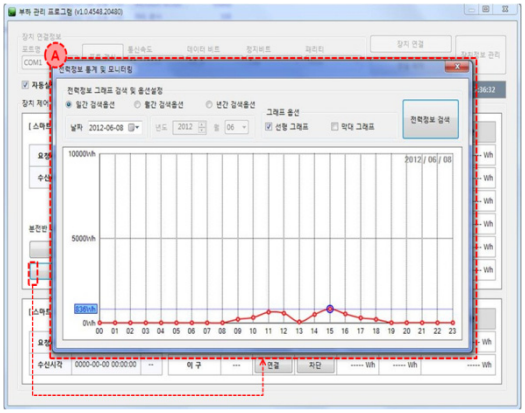


Fig. 5. Electricity information statistics and monitoring display

Click the "Accumulated Power Consumption Status" button of the smart cabinet panel in the main screen to start "Power Information Statistics and Monitoring." "Power Information Statistics and Monitoring" is shown in Area A. Search for power information by selecting "Daily", "Monthly", or "Yearly" category on the top side of Area A and selecting the date. The result of search is displayed in the graph on the bottom of the screen and the data can be shown in different graphs through graph options.

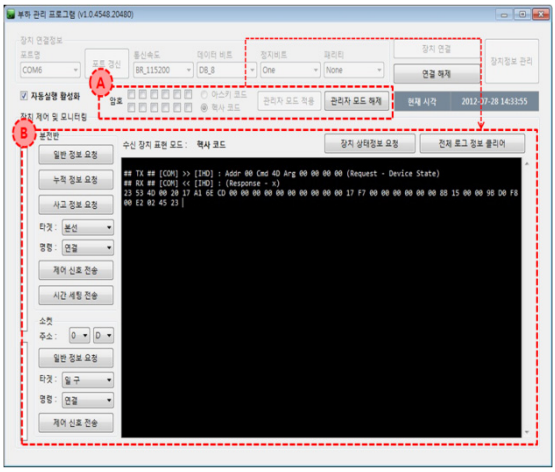


Fig. 6. Manager mode display

Enter password on the left side of Area A of the main screen and click "Apply Administrator Mode" button to start the administrator mode. The administrator mode is shown in Area B. The left side of Area B allows the administrator to arbitrarily request for certain information. The right side of the screen shows the data sent and received through UART communication.

4 Conclusion

As the Smart grid is intelligent power grid, combining information Technology to the existing power grid. Electricity suppliers and consumers exchange real-time information to two-way and is a next-generation power grid to optimize energy efficiency. This paper suggests the implementation of load management application system in energy management service environment.

In this paper, the load management system can be generally divided into 4 units: smart cabinet panel, smart socket, data receiver, and load management program. The smart socket and smart cabinet panel send and receive data through Zigbee communication and the data receiver and load management program, UART communication to monitor the information coming from smart socket and smart cabinet panel and control the device. The load management program is composed of main monitoring and control, device information management, power information analysis, accident information analysis, administrator control, XML environment setting, and database.

References

1. Lee, J.W., Park, N.: Individual Information Protection in Smart Grid. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 153–159. Springer, Heidelberg (2012)

2. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
3. Park, N.: Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
4. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
5. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)
6. Park, N., Kim, Y.: Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS, vol. 6422, pp. 193–202. Springer, Heidelberg (2010)
7. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS, vol. 6422, pp. 142–152. Springer, Heidelberg (2010)
8. Park, N.: Customized Healthcare Infrastructure using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
9. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods for Knowledge Management and Communication. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
10. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
11. Park, N.: The Implementation of Open Embedded S/W Platform for Secure Mobile RFID Reader. The Journal of Korea Information and Communications Society 35(5), 785–793 (2010)
12. Kim, Y., Park, N.: Development and Application of STEAM Teaching Model Based on the Rube Goldberg's Invention. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 693–698. Springer, Heidelberg (2012)
13. Park, N., Cho, S., Kim, B.-D., Lee, B., Won, D.: Security Enhancement of User Authentication Scheme Using IVEF in Vessel Traffic Service System. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 699–705. Springer, Heidelberg (2012)
14. Kim, K., Kim, B.-D., Lee, B., Park, N.: Design and Implementation of IVEF Protocol Using Wireless Communication on Android Mobile Platform. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 94–100. Springer, Heidelberg (2012)
15. Ko, Y., An, J., Park, N.: Development of Computer, Math, Art Convergence Education Lesson Plans Based on Smart Grid Technology. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 109–114. Springer, Heidelberg (2012)
16. Kim, Y., Park, N.: The Effect of STEAM Education on Elementary School Student's Creativity Improvement. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 115–121. Springer, Heidelberg (2012)

Toward a Mobile Application for Social Sharing Context

Meng-Yen Hsieh[1], Ching-Hung Yeh[2], Yin-Te Tsai[3], and Kuan-Ching Li[1]

[1] Dept. of Computer Science and Information Engineering, Providence University, Taiwan
{mengyen,kuancli}@pu.edu.tw
[2] Dept. of Computer Science and Information Engineering, Far East University, Taiwan
chyeh@cc.feu.edu.tw
[3] Dept. of Computer Science and Communication Engineering, Providence University, Taiwan
yttsai@pu.edu.tw

Abstract. Due to wireless and sensing technologies powerful in smartphone, a number of smartphone applications, a.k.a APPs. have combined social sharing mechanisms. This paper defines social sharing contexts on a social framework suitable to APPs. A tourism APP based on the sharing mechanisms is implemented to include social behavior for data sharing, while smartphone have supported various wireless technologies. Besides, smartphone as a hand-held device is hold by users so that various gestures with user hands are adaptive to handle the key features of sharing data between smartphones.

Keywords: social sharing, peer-to-peer, social network media.

1 Introduction

In recent years, smartphone has advanced people's habits of social sharing. Through smartphone, people share data with their friends by peer-to-peer approaches or announce data on Internet-based social media. Most social media have announced smarphone APPs for mobile users. However, smartphone still offer little room for an expensive and clean user interface where only few data is displayed at a time and scrolling becomes a chore. Currently, users share social data not only by the APPs of social network media, but also by social peer-to-peer APPs with wireless techniques.

Most smartphones have supported possible wireless techniques including NFC, Wi-Fi, and Bluetooth to allow pairing connection and communication. The Bluetooth is the popular connection for phone and its peripheral devices more early than the others. The upgrade of Bluetooth version has a high transmission rate. NFC (Near Field Communication) is one kind of wireless technique with a short distance within 10cm at rates ranging from 106 Kbps to 424 Kbps. One NFC device can sense the data of a close NFC tag with no physical touch, and transmit data with another NFC device after pairing. Wi-Fi is a popular technology for wireless local area network to support a high-speed rate of data transfer upon 150Mbps, but with the power consumption more large than the others. Wi-Fi Direct is the P2P mode of Wi-Fi.

This paper gives the social approaches of data sharing, while users assume smartphone as the major instrument to carry out social activities. In order to combine

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_15,

our past research [1, 2] where we have developed the APPs for tourism, an advanced version with social sharing mechanisms is proposed to apply user behavior of data sharing on smartphone. This paper improves the travel APP running to support social mechanisms of data sharing. The rest of this paper is organized as follows. Section 2 describes the researches related to social sharing and the APPs that we developed in the past. The context of social sharing is addressed with smartphone as the major device to share data among users in Sec. 3. Sec 4 gives the analysis results and an implementation of the revised APPs to achieve share data during traveling. We conclude the paper in Section 5.

2 Related Works

Practically all modern smartphones have more than enough hardware capability to execute software as well as that in personal computer. Some frameworks [1-3] have proposed to generate social and searching context on smartphone, and to overcome the flaws in mobile features. In fact, we have researched the establishment of mobile APPs with the two research issues. The first issue [2] is to establish the design concept of mobile APPs with NFC communications for group trip. Smartphone with NFC can simulate the operations of NFC modes as NFC Writer, Reader, and Card or Tag. The NFC design assists the activities of group trip in the interactive behavior of tourist and tour guide during a trip. The other research [1] offers a framework to integrate mobile APPs with web services releasing open APIs or resource. The mobile client includes the four components to filter data from Web services, and keep the useful information in the repository of smartphone for the moment. This framework guarantees that mobile APPs can offer data to users, even if connecting the Internet is failed.

A. Bujari *et. al*. [4] propose a proximity based social application, combining with the interesting features of the meeting and dating applications. A user can find neighbors in close distance based on his location using its geo-location service. The application offers a server to keep profiles for all users, and each user obtains the similarity scores of the detected neighbors by comparing his profile with theirs. Besides, neighboring users can create a virtual social meeting. The research [5] proposes a gesture recognition framework based on accelerometer sensor and enables seamless integration with desktop applications. The gesture information is required from accelerometer sensor embedded in smartphone device held in user's hand.

3 Social Sharing Context

3.1 Social Components

Social sharing actions are divided into two types, P2P and SN media, while smartphone is the major the instrument to share data, depicted in Fig. 1. Any two neighboring smartphones are suitable to perform P2P social sharing in order to exchange data between users; moreover, the users usually are organized in a

face-to-face social community to share files. Before sharing data, a pairing action between smartphones must be established. Two smartphones establish P2P communication through NFC or manual button as a pairing trigger for Bluetooth or Wi-Fi Direct. Sometimes the pair from three or more smartphones of a cluster is automatically established by different gestures of users holding smartphones to generate sending and receiving actions. For the reason, a cluster or pairing server is necessary for automatic pairing by user's gestures. In addition, user gestures could kindle smartphone in not only pairing, but also data delivery. Since a number of popular social network (SN) media on Internet have announced open APIs, other applications can access their data or the released Web services. Mobile APPs often achieve social actions from smartphone to SN media within the two approaches: one is that APPs deliver individual data to SN media for public announcement; the other is to capture external social data from SN for display on smartphone. Since there are various SN media on Internet for diverse business, each of them always announce open APIs or Web services with particular authentication mechanisms, individually. Therefore, to develop social APPs needs to integrate those APIs or services of the SN media for social data announcement or obtainment.

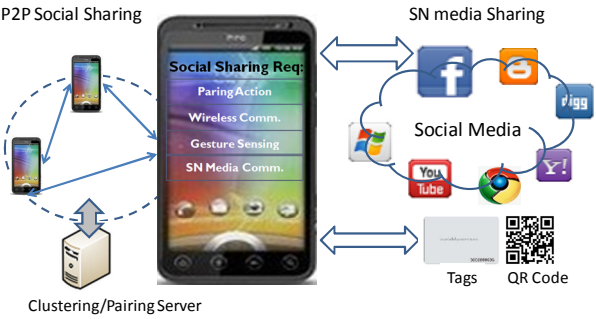


Fig. 1. Social Components for data sharing by smartphone

Four modules for social APPs are required. The first module, Pairing Action, is required to establish a pair of neighboring smartphones for P2P social sharing. The second module, Wireless Comm., chooses a suitable wireless technology for data transmission according to user's context such as user gestures or button pressing. The module of Gesture Sensing is required to determine the motion of smartphone holding by users by its sensor nodes. For example, G-sensor can detect the acceleration and direction of x-, y-, and z- axis on the physical device. The gestures for sender and receiver should be matched up very nicely, while developers take care of social actions with user gestures on device pairing and data transmission. The final module is designed to contact with SN media on Internet. To simplify the variety of authentication mechanisms and to combine released services of SN media with APPs are important for social sharing. For example, sharing data from smartphone to two or more SN media should be accomplished only in a single operation by users. Some social APPs implement social sharing actions on accessing NFC tags or QRCodes. They are an indirect way that users share data each other asynchronously.

3.2 Design Challenges for P2P Social Sharing

To design the APPs of P2P social sharing has some challenges. The first challenge is about the limitation of data size on delivery information between two neighboring smartphones. Smartphone has supported various wireless technologies with different transmission abilities. However, data with different size should be transferred with applicable wireless techniques to reduce energy consumption. Text message with very small size can be delivered by NFC, while NFC is triggering for pairing. Bluetooth is more suitable than the others to deliver a single file with a small size such an image file in a short distance. To deliver a small file using Bluetooth could has the power consumption lower than that using Wi-Fi. Transmission failure usually happens when a smartphone delivers multiple files or a single file with a big size to another smartphone by Bluetooth. Wi-Fi is applicable to the transmission with a high speed between two smartphones in a distance longer than the others.

The second challenge is how to make a pair of the two neighboring smartphones through social actions interesting to users before data transmission. The operations to directly pair two smartphones can be trigged by NFC. By the NFC context, a non-touching but closing action is necessary while users hold the sending smartphone and the receiving smartphone back to back to generate a contact of NFC detection area. Using Bluetooth and Wi-Fi techniques, social APPs must provide a touch component on screen such as button by that the sender can select manually who is the receiver. Moreover, smartphone is equipped sensors to detect physical parameters of itself and the environment where it located. To design the physical behavior of smartphone is one of major added values for social APPs interesting to users on sharing data.

4 Analysis and Implementation

4.1 Questionnaire Report

A number of social actions and user gestures are addressed, and we wanted to investigate if the users agree, disagree, or have no opinion to make social actions with the assigned user gestures. A total of 50 responses to the online survey. Table 1 represents how general users feel they could accept the patterns of user gestures that we proposed for data delivery, neighbor searching, and smartphone pairing. According the questionnaire results, most of the users agree or have no opinion on delivering a single file such as image by swinging their smartphones. More than half of the users that accepted the investigation did not agree on delivering multiple files by the user gestures of swinging or sharing. Some of them thought that to deliver multiple files with the gestures cannot flexibly determine what file should be delivered during swinging or shaking. The others generally thought that to deliver a large number of data must spend much time, thus it was not convenient to hold a smartphone with any gesture for long-time delivery. The gesture of shaking is more suitable than the rotation on smartphone. Most people are interested in applying user gestures to make any two smartphones to be a pair for communication.

Table 1. Questionnaire Details

Action Category	Items of User Gestures (Send/Recv)	Items of Statistics
Deliver a file (text or image)	Swing left/right Swing up/down Shake	Agree(39%), No-Op (39%), Disagree(22%) Agree(32%), No-Op (24%), Disagree(44%) Agree(15%), No-Op (39%), Disagree(46%)
Deliver multiple files (Album)	Swing left/right with three times Swing up/down with three times Shake with three times	Agree(24%), No-Op (37%), Disagree(39%) Agree(27%), No-Op (21%), Disagree(52%) Agree(10%), No-Op (39%), Disagree(51%)
Searching neighbors	Shake Rotation	Agree(46%), No-Op (34%), Disagree(20%) Agree(10%), No-Op (34%), Disagree(56%)
Making a pair	Make a touching Make with a circle/cross	Agree(63%), No-Op (24%), Disagree(12%) Agree(42%), No-Op (29%), Disagree(29%)

4.2 Implementation Experience

This research upgraded the APP proposed in [1] with the extended functions for social P2P and SN media communications. Fig. 2(a) gives the main screens to present the APP, consisting of MyTrip, SpotSearch, P2P, SNMedia, Alarm, and Camera. Users can plan the schedule of a personal trip by MyTrip to decide where they will go, and how long they will stay in each place. The SportSearch assists users in searching attractions, and displaying them on the Google Map. The Alarm is a warning clock to remind users of the schedule. The Camera is used to take a photo for ting with one of the tours that had been created by users before. The sub-function, friend, in Fig 2(b) assists users in searching neighbor friends by the gesture of shaking their smartphones. Users can assign the photos that they have taken into any of the tour schedules using the TripAlbum function. The P2P offers two sub-functions, NFC Hand Over and NFC Read Tag. Users can use the function, NFC Read Tag, to read a NFC tag, and to display the data on the screen. Fig 2(d) depicts the setup screen to connect with SN media. The data of a tour combines a XML file with some photos.

The APP can share individual tours on SN media and P2P modes. For social sharing on SN media, users decide what SN media are able to accept their tour data by the configuration. In advance, the APP was designed to follow the free APIs of those SN media releasing the mechanisms by that external APPs can announce the files of text and image to their discussion boards such as Blog or the wall. The APP will convert the XML data into a text file with the format corresponding to the requirement of each of the SN media with a number of image files from the access path of the pictures. For the P2P social sharing, the NFC Hand Over allows that one user shares the data of a tour with another on the P2P mode. NFC is adapted to pair two neighbor smartphones, and Bluetooth and Wi-Fi Direct are for data delivery between them. However, NFC also is responsible to data delivery, while the data of a tour is only texts of the XML file without photos. The Wi-Fi Direct is the default setting to deliver the tour data with a number of photos. While the data size of a tour with a XML file and a number of photos is less than 3M bytes or the smartphone is not supported the Wi-Fi Direct connection, the APP selects Bluetooth for the delivery.

(a) (b) (c) (d)

Fig. 2. Main screens of the travel APP supporting social sharing

5 Conclusion

This research has discussed social contexts on sharing data among users, while the users always hold smartphones to accomplish social sharing activities. The context of social sharing on developing the APPs has the challenges of SN media integration, and device pairing and data transmission on P2P. A questionnaire for P2P social sharing is offered to improve the acceptability of gesture design on data transmission between smartphones to users. According to the analysis on questionnaire findings, the APP for trip planning that we developed has been upgraded for social sharing.

Acknowledgements. This investigation was supported by National Science Council, Taiwan, under grants NSC 101-2221-E-126-006-, NSC 101-2221-E-126-002-, and NSC 99-2632-E-126-001-MY3.

References

1. Hsieh, M.Y., Lin, H.Y., Yeh, C.H., Li, K.C., Wu, B.S.: A Mobile Application Framework for Rapid Integration of Ubiquitous Web Services. In: 9th IEEE International Conference on Ubiquitous Intelligence and Computing, Japan (September 2012)
2. Hsieh, M.Y., Wu, T.Y., Tsai, Y.T., Shih, C.H., Li, K.C.: Interactive Design Using Non-Touch Technologies for Group Trip. In: The 2012 IEEE International Symposium on Intelligent Signal Proc. and Comm. Systems (November 2012)
3. Konstantinidis, A., Aplitsiotis, C., Zeinalipour-Yazti, D.: SmartP2P: A Multi-objective Framework for Finding Social Content in P2P Smartphone Networks. In: Proc. MDM, pp. 324–327 (2012)
4. Bujari, A., Miotto, N.: Nudge Nudge: a Proximity Based Social Application. In: Wireless Days (WD). IFIP, pp. 1–3 (2011)
5. Rajanna, V.D.: Framework for Accelerometer Based Gesture Recognition and Seamless Integration with Desktop Applications. International Journal of Scientific and Research Publications 3(1) (2013)

A Research Based on the Effect of Smart Phone Use on Consumption Life of Teenagers in a Smart Era

Jeonghan Son[1], Keon Uk Kim[2], Yeon-gyeong Seo[3], Wonyeong Oh[4], Seowon Choi[5], and Ana Kang[6,*]

[1] Daejeon Mannyon High School
[2] Daejeon Jungang High School
[3] Chungnam Girls'High School
[4] Jungil High School
[5] Daejeon Dunsan Girls'High School
[6] Engineering Education Innovation Center, Mokwon University
anakang37@gmail.com

Abstract. An advent of the smart era, including a smart phone, tablet, PC, online community, etc, brings a groundbreaking change in everyday life. Internet connection is possible at any time. Also, smart era, including smart home, smart building, smart city, etc, is beyond a simple technology, and brings an enormous change in a way of communication in daily life and even economic life. This smart era also brings a big change in the life of teenagers. This research points out that the economic value consumed in return for the convenience by teenagers in the smart era is increased, and their consumption becomes unwise and unreasonable. This research finds out the problem of teenagers' consumption activity in the smart era and proposes a solution to improve the problem.

Keywords: Smartphone, consumption, teenager.

1 Introduction

Smart phone is beginning to be accepted as a part of teenager's life. According to smart phone using state survey of communications commission and Korea Internet Security Agency in December, 2010, compared to May, 2010, the rate of new smart phone users between the ages of 12 and 19 is increased from 9.5% (May) to 15.3%(December) in December, 2010. Due to supply of the smart phone, teenagers can connect to internet for 24 hours, anywhere and at any time, find the an information, play a network game, watch a movie, and listen to an online lecture. These activities profoundly change the lives of teenagers. Therefore, teenagers' leisure life and way of life rapidly become convenient, and the economic value consumed in return for convenience by teenagers is increased. Teenagers' consumption becomes unwise and unreasonable. This research tries to find out the problem of teenagers' consumption activity and propose a solution to improve the problem [1].

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_16,

This research is composed of 4 chapters. The second chapter is about the smart era and smartphone in related research, and the economy of teenagers and unwise consumption. The third chapter is about the effect of smartphone use on consumption life of teenagers. The last chapter is about a conclusion of the research.

2 Related Works

2.1 Smart Era and Smartphone

The smart era is period of introduction of ubiquitous age, a period in which the possibility of new economic activities, including UCC, blog, etc, is revealed, and a society where smart home, smart building, and smart city are possible. Smart phone is ranked first in teenager IT major trend section of a report called "vision and policy direction of 2011 IT" by Korean Internet and Security Agency. This reveals the importance of smart phone in technology of information and communication and in everyday life. Smart phone can be defined as "a movable multi-media which has a high level of UI expandability through existing cellphone's OS(Operating System) and is capable of running diverse and personalized application". Due to a rapid expansion of smart phone use, there is a concern about excessive use of smart phone by users. The applications can be downloaded from online, and these applications can be used while offline. This is the characteristic of the smart phone. This specialized function of smart phone provides a specialized convenience which cannot be found in the existing cellphone, and this convenience might cause the addiction of the smart phone. The research up to now is about future life, the economic effect, and technological trend. No research is written in a perspective or recipient [1-9].

2.2 Economy of Teenagers

On this point, we want to focus on making a point that the economic life means more than a mere concept of materials. Efficient economic activities are the activities that get the greatest satisfaction, and make the largest profits as possible from the regularly provided resources. Youth economy, according to Mankiw's definition, can be seen as a reasonable consumption to the youth. They have to make a decision at the provided range of income, and time. Action of consumption which doesn't only consider present, but also considers further future in order to maximize the satisfaction is the main point of this conceptualized youth economy on this paper. In other words, economy, to the youth, is the action of using the provided finance resources, and time in a most reasonable way or is simply the action of consumption [10].

2.3 Making an Unwise Consumption of Teenagers Living in Smart Era

The advent of cyber leisure activity based on information technology interrupts a reasonable consumption and proper management of time. Expense of teenagers' cyber

leisure activity is maximum average 250,000 a month which is more than an average allowance of teenagers. The expense of cyber leisure activity is mostly made by using the cell phone and wire telephone. Through cell phone and wire telephone, the payment can be easily made. This shows that the problem of teenagers' consumption become worse. Also, since online game is highly addictive, interruption of everyday life, illegal access to personal information, default on payment, and plunder of item for the purchase of item become social problems. Smart era provides more convenient life to teenagers. At the same time, it contain a danger of making teenager make an uneconomic consumption. A reasonable consumption means a proper selection of the several services and goods in the range of given income and time, and consumption behavior which maximizes the satisfaction by considering the present and the distant future. However, consumption standard of teenagers is not settled in smart era which rapidly enters into the everyday life. Teenagers'' obsession with smart phone causes an interruption of everyday life and school work, intensification of game addiction, and raise in expense of cellphone usage and wireless fee are consumption problems of teenagers. This research calls a concept of teenagers' consumption as "making a fool of" which is opposite concept of the smart era. Teenagers do not realize their own problems of consumption, their lives are controlled by smart era, and the problems of consumption problems become worse [11-13].

3 The Effect of Smartphone Use on Consumption Life of Teenagers in the Smart Era

With a hypothesis "students who use the smart phone are likely to incur a danger of making an unwise consumption than the students who do not use the smartphone", this chapter tries to find out the effect of smartphone use on the consumption life of teenagers through a questionnaire design and analysis of questionnaire result.

3.1 Questionnaire Design

To verify the hypothesis that unwise consumption of teenager who use the smartphone becomes worse, the questionnaire is conducted with high school students as subjects. The purpose of questionnaire is to prove a difference of consumption life between smartphone users and smartphone nonusers. This will prove a hypothesis that students who use the smart phone have a higher possibility of having an unwise consumption than students who do not use the smart phone. To serve the purpose of questionnaire, a variable for setting a control group, a question intended for smart phone user, the question intended for existing cellphone user, cellphone user and smart phone nonuser, economic activity, and the article about internal review for verification of questionnaire are considered. The article which is not considered is deduced by conducting a pre-questionnaire for some high school students [14].

3.2 A Result of Questionnaire and Analysis

A final questionnaire survey is based on 8 high schools located at Seoul, Daejeon, and Jeonju and conducted from June 11 to June 25(about two weeks). Total 300 students are surveyed. Among 300 students, 150 students possess the smart phone, and 150 students possess the regular cellphone.

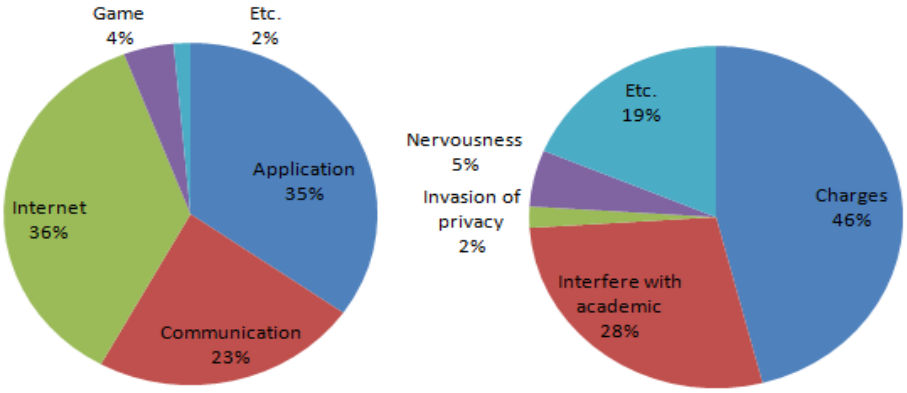


Fig. 1. Convenience and inconvenience of using a smart phone

- **Making an Unwise Economy of Teenagers – Monetary Aspect**

The number of smart phone user who spend more than seventy thousand won is about eight times more than the number of regular phone user who spend more than seventy thousand won. More than 100 regular phone user pays a less than thirty thousand won for a wireless fee. A lot of smart phone users spend between thirty thousand and seventy thousand won for the wireless fee. Considerable number of students spends more than seventy thousand won for the wireless fee. The expense of the allowance and total wireless fee of smart phone user is high. There is also considerable number of smart phone user who uses a supplementary service which costs between five thousand won and twenty thousand won. Besides, there are 15 respondents who spent more than thirty thousand won for the supplementary service fee. As teenagers use the smartphone, there are many more teenagers who spend money on supplementary service than the teenagers who do not use the smartphone. The research tries to figure out whether the teenagers have an experience of having a part-time job to pay the smart phone fee. As a result of survey, there is a small number of respondents who have an experience of having the part-time job among the smart phone users and regular phone users. Examining the rate of the result, 4.7% of the regular phone users have an experience of working while 11.3% of the smart phone users have the experience of working. 11.3% is more than half of 4.7%.

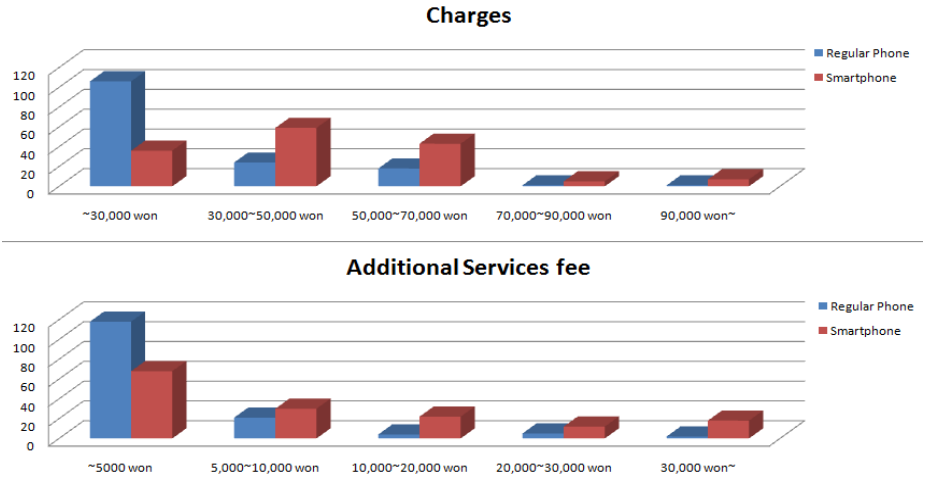


Fig. 2. Comparison Charges and additional services fee

- **Making an Unwise Economy of Teenagers – Time Aspect**

To examine a time aspect of making an unwise economy of teenager, the method of asking actual using hours and habits is chosen. The questions about using hours is asked to know the frequency of use, and about using habits is asked to figure out the degree of interest in cellphone and smart phone. For questions asking for the using hours, most cellphone and smartphone users answer that they use their phone anytime they want. Smart phone users have a higher rate of using the phone freely than smart phone nonusers. On the other hand, for the using habits, the cases of serious addiction symptom – using the phone all day as a part of life and constantly checking the phone – are similar for smart phone users and nonusers. There is a higher number of smart phone users who shows the addiction symptom by using the phone for most of every-day life and checking the phone every 10~20 minutes than the regular phone user.

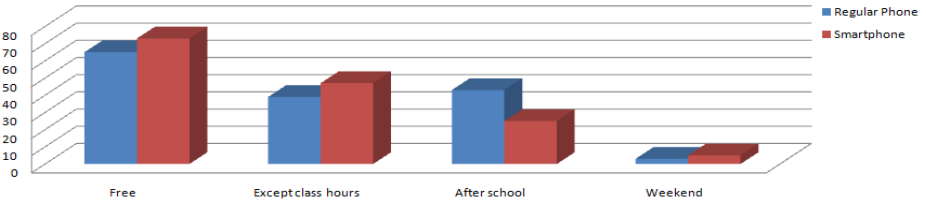


Fig. 3. Comparison of time to use the regular phone and smartphone

4 Conclusion

This research tries to prove a difference in consumption life between smart phone users and nonusers through a questionnaire survey based on 300 students. It also tries

to prove that students who have a higher frequency use of smartphone tend to have a higher possibility of having an unwise economy. Unwise consumption appeared in unwise economy remarkably appears in monetary aspect than in time aspect. Also, smartphone interrupts the study time. Since students have a part time job to pay the fee, time becomes an opportunity cost for the pleasure. Eventually, smart phone users spend their money on expensive fee and many hours as economical expense. These results showed that teenagers do not recognize the problems of their own consumption, their lives are controlled by smart era, and their problems of consumption become worse.

To teenagers, smart phone has a more meaning than just a way of making an active communication. Therefore, "smart consumption' is needed to teenagers who have undeveloped economic viewpoint in this smart era. For smart consumption, a reasonable way of thinking is required. For this, 'economic efficiency' has to be achieved. For this, two aspects of expense and utility should be considered. In an aspect of expense, teenagers should reduce the wasted expense of smart phone. Writing an account book will lead to a regulation of one's expense according to one's circumstances and decrease the expense. In an aspect of utility, teenagers should analyze the factors that give satisfaction to them. Eventually, a solution is that they should recognize the smart consumption which fits for the smart era and act properly.

References

1. Smartphone use survey (II), KCC/KISA (2010)
2. Gwon, G., Im, T., Choi, W., Park, S., Oh, D.: Future smartphones that opens. CEO Information 741 (2010)
3. 2011 IT trends, forecasts and policy direction, NIA (2011)
4. Kim, D., Tai, J.: A Study on the Mediation Experiences of Smart Phone Users. The Content of Humanities 19, 373–394 (2010)
5. Republic of Korea Mobile Almanac 2009, KISA (2009)
6. Kim, B.: Smart phones, mobile ecosystem change, Dong-A Ilbo (February 11, 2010)
7. Kang, J.: Holding a smart phone in the toilet, or you addicted? The Korea Economic Daily (July 22, 2010)
8. No, M., Kim, J., Lee, J.: An Exploratory Study on Smart-Phone and Service Convergence. Society for e-business Studies 15(4), 59–77 (2010)
9. Hwang, H., Son, S., Choi, Y.J.: Exploring Factors Affecting Smart-Phone Addiction - Characteristics of Users and Functional Attributes. Korean Journal of Broadcasting 25(2), 277–313 (2011)
10. Mankiw, Mankiw's economics, Kyobo Book, 205
11. Park, M.J.: Analysis of Consumption Behavior of the Youth in Cyber, Leisure Activities. MS Thesis, Ewha Woman's University (2009)
12. National Council of Homemakers' Classes Daejeon Branch, young people use online games and Myspace Survey. A Monthly Consumer 301, 10–12 (2008)
13. Han, J.: 200 intellectuals to think of Adolescent Health Declaration. The Korea Economic Daily (March 8, 2011)
14. Jeong, Y.: Chapter shutdown, internet games Year 'mobile' applies even one. The Kyunghyang Shinmun (March 28, 2011)

Quality-Workload Tradeoff in Pig Activity Monitoring Application

Haelyeon Kim, Yeonwoo Chung, Sungju Lee, Yongwha Chung, and Daihee Park

Dept. of Computer and Information Science,
Korea University, Sejong, Korea
{mahakhl,william0516,peacfeel,ychungy,dhpark}@korea.ac.kr

Abstract. Generally, there is a tradeoff between quality and computational workload required to obtain that quality. In this paper, we focus on practical issues in implementing a pig activity monitoring system. We first propose a method for evaluating the quality-workload tradeoff in the activity monitoring application. Then, we derive the cost-effective solution within the acceptable range of quality for the activity monitoring application. Based on the experiments with the video monitoring data obtained from a pig farm, our method can derive the cost-effective resolution size and frame rate without degrading the accuracy significantly.

Keywords: Activity Monitoring, Quality, Accuracy, Workload, Tradeoff.

1 Introduction

Early detection of anomalies is an important issue in the management of group-housed livestock. In particular, the damage caused by the recent outbreak of livestock diseases in Korea such as foot-and-mouth disease was serious. In order to minimize the damage incurred from such diseases, it is necessary to develop the technology for collecting and analyzing livestock activity data. Although some progress in monitoring livestock has been made recently in Korea[1-2], practical issues in implementing an automated activity monitoring system with a video sensor have not yet been reported. This is also true in other countries, although some research studies on applying Information Technology(IT) to livestock management have been reported in the last decades[3-8].

In this paper, we apply the computer vision technology to the daily-life activity monitoring of group-housed pigs in order to manage the healthcare problem of pigs[9]. Especially, caring individual pigs by farm workers in a large-scale pig farm is almost impossible. For example, a pig farm where we obtained the video monitoring data has more than 20000 pigs and 1000 pig rooms. Caring these pigs with 10 farm workers is almost impossible, and an automated analysis of the daily-life activity is required. Furthermore, we should consider the practical issues such as cost in implementing the automated activity monitoring system because poor profitability of pig farming has inhibited large-scale investment. In addition to this initial investment,

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_17,

managing individual devices such as pedometers[10] in a large-scale pig farm may not be acceptable because of the managing cost. Thus, we consider a video camera which does not need such managing overhead once installed, in order to monitor the group activity of pigs.

In a video-based monitoring system, we first extract the activity data from the video, and then determine the anomaly based on the amount of activities extracted. The details of the activity monitoring system are shown in Fig. 1. However, simple methods for extracting the daily-life activity such as the frame difference method[11] may not guarantee the required accuracy. In order to provide the required accuracy, more complicated methods such as Gaussian Mixture Model(GMM)[11] have been widely used in video-based monitoring systems. In this paper, we also apply GMM in implementing a monitoring system, and analyze the tradeoff between the quality and the computational workload for implementing a practical system.

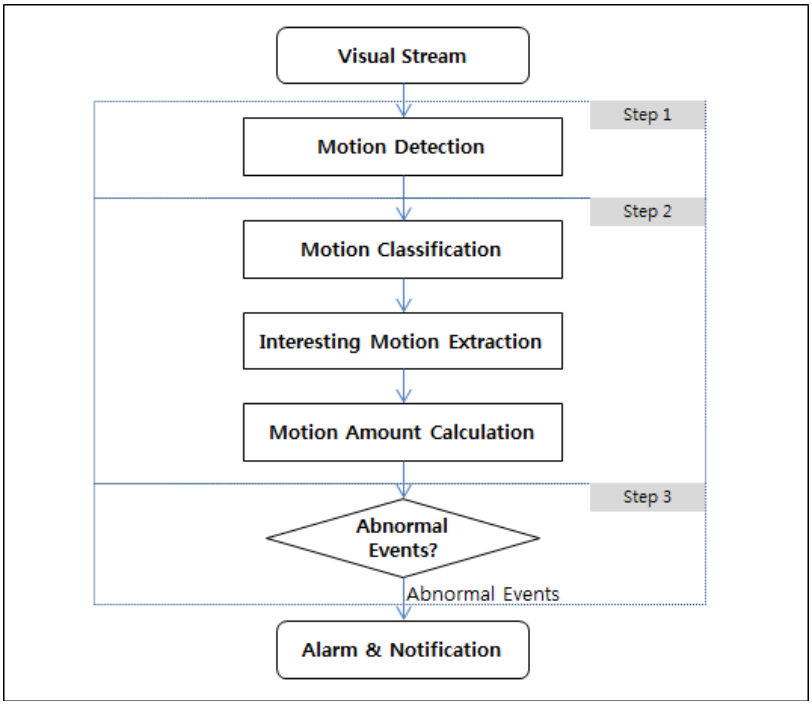


Fig. 1. Flowchart of activity monitoring system

In fact, a similar tradeoff issue has been widely studied in the video compression community recently[12-13]. That is, this upcoming compression technique tries to maintain the required Peak Signal to Noise Ratio(PSNR), a quality metric widely used in the video compression community, with a minimum workload.

In the activity monitoring applications for livestock, however, a reasonable metric for quality has not been reported. Thus, we first propose a quality metric(*i.e.*, accuracy)

which can be applied to the activity monitoring applications. Then, we can derive the minimum resolution size and frame rate for reasonable(*i.e.*, close to the baseline case with the maximum resolution size and frame rate) accuracy. To the best of our knowledge, this is the first report of the tradeoff in monitoring the continuous and large incoming data stream that is a characteristic of monitoring systems, and we verify our solution with real video data acquired from group-housed pigs. Although we analyze the tradeoff for obtaining a cost-effective solution, the proposed method can also be applied to activity monitoring applications with portable devices/battery-operated sensors in order to prolong the battery lifetime.

This paper is organized as follows. Section 2 describes the proposed activity monitoring system, and the details of the implementation and experimental results are explained in Section 3. Finally, we provide some concluding remarks in Section 4.

2 Pig Activity Monitoring

In this paper, we focus only on the automated activity monitoring system for weaning pigs. In fact, this system can be integrated with the existing control system which controls illumination, temperature, humidity, CO_2, etc, and sends an alarm in case of emergency. From a camera installed at the ceiling of a pig room, the 24 hours/365 days visual stream data are transmitted to a server through a LAN cable.

The system determines first whether a given scene has any motion or not. Note that, the simplest pixel difference algorithm in order to detect any possible motion in a given scene is known to be inaccurate[11]. Therefore, we need to apply more complex algorithm such as GMM for more accurate monitoring. However, applying GMM straightforwardly to the 24 hours/365 days visual stream data generated from a large-scale pig farm may require too much implementation cost. That is, we need to find the optimum tradeoff between the accuracy and the computational workload. Once we find this optimum tradeoff, we can adjust the camera setting or downsample the input data.

For finding the optimum tradeoff, we first set the resolution size and frame rate to the maximum value supported by the camera(*i.e.,* called *base-case*), and then compute an hourly activity data obtained from the activity monitoring data. After setting the resolution size and frame rate to each downsampled value(*i.e.,* called *downsampled-case*), we compute another hourly activity data obtained from the downsampled activity monitoring data. Then, we compute the similarity between the base-case and the downsampled-case.

Generally, the correlation of two random variables is a standard measure of how strongly two variables are linearly related. Correlation therefore naturally captures our intuitive notion of temporal similarity. Note that, the temporal similarity between the two cases(*i.e.*, the base-case and the downsampled-case) can be computed by

$$\frac{1}{T}\sum_i \left(\frac{X_{b,i}-\mu(X_b)}{\sigma(X_b)}\right)\left(\frac{X_{d,i}-\mu(X_d)}{\sigma(X_d)}\right) \quad (1)$$

Let $X_{b,i}$ be an hourly activity data obtained from the base-case at each hour i and $X_{d,i}$ be an hourly activity data obtained from the downsampled-case at each hour i. Note that, an hourly activity data means the accumulated amount of activities detected by GMM during an hour. For the base-case b, $\mu(X_b)$ and $\sigma(X_b)$ denote the mean and the standard deviation, respectively. For a particular downsampled-case d, $\mu(X_d)$ and $\sigma(X_d)$ denote the mean and the standard deviation, respectively.

Finally, we measure the execution time and the power consumption separately, and decide which one is better for the required workload.

3 Experimental Results

In this section, we present some experimental results to derive the optimal resolution size and the optimal frame rate under the reasonable accuracy. The activity monitoring system was comprised of a video sensor and a server. Note that, we installed a HD camera at the ceiling of a pig room and set the resolution size to 1280× 720 pixels and the frame rate to 30 frames/second(fps) initially(*i.e.,base-case*). Feed and water were available ad libitum, and light was provided continuously. With this initial setting, we measured the amount of activities caused by 19 pigs in a pig room. The experiments were performed on an Intel Core i5-2500 at 3.3GHz 4-core processor with 4GB RAM, and we also measured the actual power consumption with WT210 Digital Power Meter tool. We have used the GMM background subtraction to detect the motion and OpenCV(Open Source Computer Vision) version 2.3 as a wrapper library to import video sequence to the C program.

From the base-case of 1280×720 pixels and 30 fps, we could obtain various downsampled resolutions 640×480, 320×240, 160×120, 80×60 pixels and downsampled frame rates 10, 15, 1 fps. Fig. 2 shows some examples of the hourly activity pattern with various resolution/frame rate cases. That is, some cases have similar patterns(*i.e.*, 320×240 pixels/10 fps *vs.* 320×240 pixels/1 fps), whereas other cases have entirely different patterns(*i.e.*, 320×240 pixels/10 fps *vs.* 80×60 pixels/10 fps). Thus, we should derive the downsampled-case whose hourly activity pattern is similar to that of the base-case with minimal computational workload.

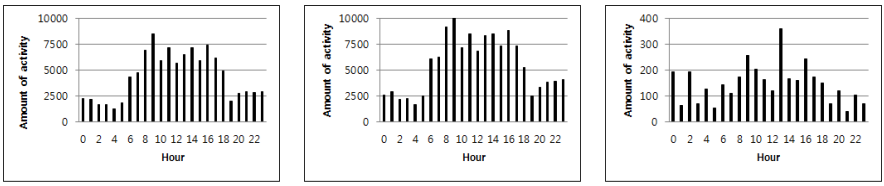


Fig. 2. The hourly activity patterns with various resolution/frame rate cases

For describing the temporal similarity between the base-case and the downsampled-case, we computed the similarity using Equation (1). Table 1 shows the similarity values normalized to the base-case(*i.e.,* the similarity of the base-case itself

is *1*). Note that, the similarity is more affected by resolution than frame rate. For describing the workload, we measured the execution time and power consumption, and then computed the energy consumption. We confirmed that the power consumption was irrelevant to various resolution/frame rate cases. Therefore, the execution time can be used as computation workload. Table 2 shows the relative execution time normalized to the base-case(*i.e.,* the relative execution time of the base-case itself is *1*). Finally, we represent the relative value of "accuracy/workload" tradeoff in Table 3(*i.e.,* the relative tradeoff of the base-case itself is *1*). Therefore, we can derive the cost-effective resolution size and frame rate as 160×120 pixels and 1 fps, with enough accuracy(*i.e.*, the similarity is 0.9).

Table 1. Similarity(*i.e.*, accuracy) with various resolution/frame rate cases

frame rate / resolution	15 fps	10 fps	5 fps	1 fps
640×480 pixels	0.99	0.96	0.95	0.95
320×240 pixels	0.96	0.95	0.94	0.93
160×120 pixels	0.92	0.90	0.90	0.90
80×60 pixels	0.34	0.56	0.64	0.69

Table 2. Relative execution time(*i.e.*, workload) with various resolution/frame rate cases

frame rate / resolution	15 fps	10 fps	5 fps	1 fps
640×480 pixels	0.24	0.07	0.05	0.04
320×240 pixels	0.10	0.04	0.04	0.03
160×120 pixels	0.06	0.03	0.03	0.03
80×60 pixels	0.05	0.03	0.03	0.03

Table 3. Relative tradeoff(=accuracy/workload) with various resolution/frame rate cases

frame rate / resolution	15 fps	10 fps	5 fps	1 fps
640×480 pixels	4.06	12.38	17.14	19.92
320×240 pixels	9.72	20.20	23.48	24.64
160×120 pixels	14.65	23.13	25.04	25.49
80×60 pixels	6.38	15.15	18.60	20.28

4 Conclusion

Automated detection of abnormal behaviors in livestock is an important issue in livestock management. We proposed a cost-effective, automated technique for analyzing weaning pig's activities using visual information acquired from a camera installed in the pig's house. Especially, this research focused on the practical issues such as the quality-workload tradeoff in implementing a pig activity monitoring system. From the experiments, we found that our method can satisfy the low cost requirement without degrading the accuracy significantly.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2043679).

References

1. Hwang, J., Yoe, H.: Study of the Ubiquitous Hog Farm System using Wireless Sensor Networks for Environmental Monitoring and Facilities Control. Sensors 10, 10752–10777 (2010)
2. Hwang, J., Shin, C., Yoe, H.: Study on an Agricultural Environment Monitoring Server System using Wireless Sensor Networks. Sensors 10, 11189–11211 (2010)
3. Berckmans, D.: Automatic On-line Monitoring of Animals by Precision Livestock Farming. In: Keynote in the ISAH Conference "Animal Production in Europe: The Way forward in a Changing World", vol. 1, pp. 27–30 (2004)
4. Cox, S. (ed.): Precision Livestock Farming. Academic Pub. (2003)
5. Davies, E.: The Application of Machine Vision to Food and Agriculture: a Review. The Imaging Science J. 57, 197–217 (2009)
6. Frost, A., Schofield, C., Beaulah, S., Mottram, T., Lines, J., Wathes, C.: A Review of Livestock Monitoring and the Need for Integrated Systems. Comput. Electron. Agric. 17, 139–159 (1997)
7. Ruiz-Garcia, L., Lunadei, L., Barreiro, P., Robla, J.: A Review of Wireless Sensor Technologies and Applications in Agriculture and Food Industry: State-of-the-art and Current Trends. Sensors 9, 4728–4750 (2009)
8. Wathes, C., Kristensen, H., Aerts, J., Berckmans, D.: Is Precision Livestock Farming an Engineer's Daydream or Nightmare, an Animal's Friend or Foe, and a Farmer's Panacea or Pitfall? Comput. Electron. Agric. 64, 2–10 (2008)
9. Ekesbo, I.: Farm Animal Behavior: Characteristics for Assessment of Health and Welfare. CAB International (2011)
10. Brehme, U., Stollberg, U., Holz, R., Schleusener, T.: ALT Pedometer — New Sensor-Aided Measurement System for Improvement in Oestrus Detection. Comput. Electron. Agric. 62, 73–80 (2008)
11. Hu, W., Tan, T., Wang, L., Maybank, S.: A Survey on Visual Surveillance of Object Motion and Behaviors. IEEE Tr. Systems, Man, and Cybernetics – Part C34, 334–352 (2004)
12. Lian, C., Chien, S., Lin, C., Tseng, P., Chen, L.: Power-Aware Multimedia: Concepts and Design Perspectives. IEEE Circuits and Systems Magazine, 26–34 (2007)
13. He, Z., Cheng, W., Chen, X.: Energy Minimization of Portable Video Communication Devices based on Power-Rate-Distortion Optimization. IEEE Tr. Circuits and Systems for Video Technology 18(5), 596–608 (2008)

Applying Different Cryptographic Algorithms for Mobile Cloud Computing

Sung-Min Jung[1], Nam-Uk Kim[1], Seung-Hyun Lee[1], Dong-Young Lee[2], and Tai-Myoung Chung[1]

[1] Dept. of Electrical and Computer Engineering, Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu, Suwon-si, Gyeonggi-do, Korea
{smjung,nukim,shlee87}@imtl.skku.ac.kr, tmchung@ece.skku.ac.kr
[2] Dept. of Information and Communication, Myongji College, 356-1 Hongeun3-Dong, Seodaemun-Gu, Seoul, Korea
dylee@mjc.ac.kr

Abstract. Cloud computing is new paradigm to use computing resources that are delivered as services over a network. It optimizes the usage of IT resources such as CPU, storage, and network. Many services related of cloud computing are popular to end users and it is becoming more important these days. There are many smart phones, smart pads and other mobile devices and end users can access to cloud computing environment through these mobile devices. Thus, they can use powerful computing resources on their physical devices. This environment indicates mobile cloud computing in this paper. There are two devices such as a physical device in real world and a virtual device in cloud computing environment. Service providers should use strong cryptographic algorithms to guarantee secure communication between a physical device and a virtual device. However, the strong cryptographic algorithms waste time to process each tasks and it causes network congestion. The network congestion occurs when a physical device is processing too many data packets. Also, it cannot be guaranteed its network quality of service. We need to consider the network quality of service to avoid this congestion. We should try to reduce the execution time to guarantee quality of service. We propose suitable method that the cryptographic algorithms with different key lengths at various environment.

Keywords: Cloud computing, QoS management, Cryptographic algorithm.

1 Introduction

According to Wikipedia, cloud computing indicates the use of computing resources that are delivered as services over the internet. Virtualization is an essential technique to cloud computing, since it integrates computing resources which are physically different location and it distributes these resources to users dynamically. In general, the service layers of cloud computing can be divided into three layers such as Software as a Service, Platform as a Service, and Infrastructure as a Service[1][2].

In this paper, we propose the suitable way to reduce the execution time to guarantee quality of service in mobile cloud computing. The execution time can be

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_18,

divided into processing time and encryption time. The processing time is the time it takes to process its tasks and encryption time is the time it takes to encrypt or decrypt data. We propose a suitable method that cryptographic algorithms with different key lengths are used in various environments. If we can use different key lengths dynamically, the encryption time would be managed efficiently and finally we can manage the execution time. Therefore, we can avoid the network congestion easily as decreasing the execution time in mobile cloud computing[6][7].

The remainder of this paper is divided into three sections. Section 2 discusses the concept of mobile cloud computing and a cryptographic algorithm. In section 3, we discuss the suitable method to guarantee the quality of service by using the different key lengths. Finally, section 4 concludes this paper.

2 Mobile Cloud Computing

There are numerous servers and network equipments in a data center and they are configured the cloud computing environment. The number of mobile devices such as smart phones and smart pads grows rapidly recently. End users can access easily to cloud computing environment thought these mobile devices[8][9]. Figure 1 shows this environment and we call it to mobile cloud computing in this paper. We define that mobile cloud computing is one of specific services of cloud computing and it is a mobile service which is added a cloud computing service. As shown the figure 1, mobile cloud computing is configured real mobile devices, infrastructure resources and virtual devices in data center. End users approach to virtual devices in cloud computing using mobile devices such as smart phones and smart pads. User's applications are in cloud computing and virtual devices provide the job outcome of applications via internet to mobile devices.

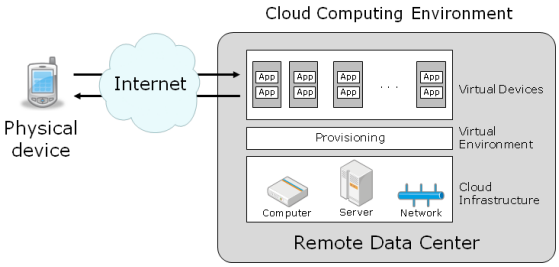


Fig. 1. Mobile cloud computing

In general, the physical devices are mobile devices and they have limited resources such as low CPU, small storage and memory capacity[3]. Therefore, if we use physical devices to access to virtual devices in cloud computing, then we should consider the possibility of network congestion which is caused from their limited resources. Also, there are many issues at a security aspect in this environment and specific requirements for cloud computing is still being discussed. Service providers

should offer strong cryptographic algorithms to guarantee secure communication between a physical device and a virtual device in cloud computing environment. However, strong cryptographic algorithm needs more time to encrypt and decrypt data packets and it would cause network congestion. There is a tradeoff between to reduce execution time and to guarantee secure communication[10]. We should consider the quality of service to avoid this congestion in mobile cloud computing and the secure communication at the same time.

3 Architecture of Mobile Device

As shown in the figure 2, we proposed the architecture of a mobile device to operate mobile cloud computing in our previous work[4]. There are input/output manager, service manager, security manager, and communication manager. However, there is no consideration of the factors related of quality of service. Thus, we revise the architecture to support the quality of service of applications in this paper.

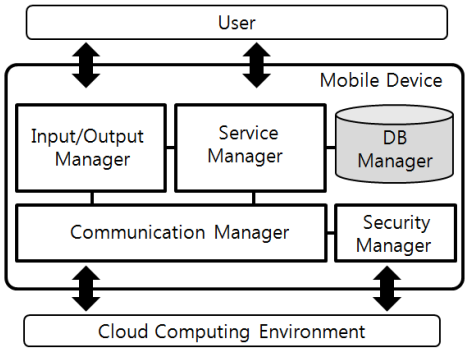


Fig. 2. Architecture of a physical device

We try to reduce execution time to support quality of service. This execution time can be divided to processing time and encryption time. We propose suitable way to use the cryptographic algorithms with different key lengths. We suggest the architecture of mobile device and we define the security manager related to keep the security requirements. The security manager is in communication manager and it handles data encryption and decryption with different key lengths. We choose the small size of cryptographic key in network congestion and we avoid this congestion easily.

Figure 3 shows the detail configuration of a security manager. It is configured a cryptographic module and an execution time checker. The cryptographic module is responsible for a local data protection and message protection. It manages the cryptographic algorithms and different key sizes. The execution time checker figures out the execution time with a specific threshold. When physical devices approach to virtual devices in cloud computing, physical devices should receive a lot of data packets from the virtual device. Therefore, this feature can cause the network congestion frequently.

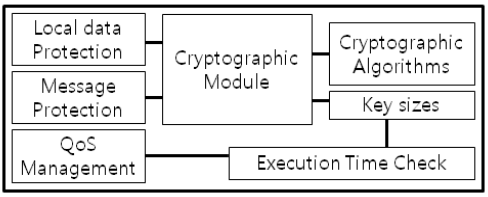


Fig. 3. Security manager

We use an encryption algorithm with different key sizes. If there is network congestion, we choose a smaller key size than the previous. Then, execution time to encrypt data can be reduced and we can avoid the network congestion. We assume that a cloud computing environment consists of a set of N tasks $< T_1, T_2, T_3, \cdots T_N >$ as shown the figure 4.

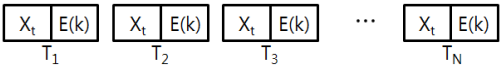


Fig. 4. Set of tasks

Equation (1) shows the execution time of a task T_i .

$$X_i = X_t + E(k) \tag{1}$$

X_t is the processing time to execute a task and $E(k)$ is the execution time to encrypt data by the k key length. It would be different according to different key sizes. The execution time meets the specific threshold and we choose suitable key size and apply it immediately in the system. The threshold may be queueing delay. Equation (2) shows the network latency of a task T_i [5].

$$Y_i = \frac{M}{B - \lambda \cdot M} \tag{2}$$

In this equation, M is the message length, B is the bandwidth and λ is arrival rate. As shown the figure 5, we define the control flow of our proposed way.

We assume that AES algorithm is used in our proposed scheme. There are three key length such as k[0]=128, k[1]=192 and k[2]=256. While physical device operate tasks from i to N, we calculate queueing delay Y_i. Then, we compare specific threshold and the sum of execution time and queueing delay. If the result is larger than specific threshold then we select the smaller key length to reduce the cost for encryption. Otherwise, if the result is smaller than specific threshold then we choose the longer key size than previous to enhance the security.

```
m = 2
k[m] = [128, 192, 256]
For i = 1 to N
    Calculate X_i
    IF Y_i  > threshold AND m > 0 THEN
        k[m] = k[m - 1]
    ELSE IF m < 2 THEN
        k[m] = k[m + 1]
    ELSE
        Calculate X_t
        Calculate E(k[m])
        X_i = X_t + E(k[m])
    ENDIF
ENDFOR
```

Fig. 5. The Execution Time

4 Conclusion

The concept of mobile cloud computing is a mobile service which is added a cloud service that using virtualization technology to computing resources such as server, storage and network. Recently, there are various smart phones, smart pads and other mobile devices and clients can access to cloud computing environment via these devices at everywhere. Since mobile cloud computing system has to deal with very large amount of data, it is difficult to satisfy this requirement under the limited resources of mobile device. Therefore, it is important to support quality of service because the mobile devices have a very restricted resources and their performance is lower than general desktops. Furthermore, cryptographic algorithms are essential to keep data authentication, integrity and confidentiality. However, these algorithms cause network congestion. There is a tradeoff between to keep the security and to provide appropriate QoS.

In this paper, we propose the enhanced way to select suitable cryptographic algorithms to reduce the encryption time. We choose a suitable key length when there is network congestion. In the future, we wish to evaluate the performance under the more realistic factors.

Acknowledgements. This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012-0005861).

References

1. Montero, R.S., Llorente, I.M., Foster, I.: Virtual Infrastructure Management in Private and Hybrid Clouds. IEEE Internet Computing 13(5), 14–22 (2009)
2. Buyya, M.R., Pandey, S., Vecchiola, C.: Cloudbus toolkit for market-oriented cloud computing. In: Proceedings 1st International Conference on Cloud Computing (CloudCom 2009), pp. 3–27 (2009)
3. Appisty Inc.: Cloud Platforms vs. Cloud Infrastructure, White Paper (2009)
4. Kim, N.-U., Eom, J.-H., Kim, S.-H., Chung, T.-M.: The Architecture of User-end Client for Mobile Clout Service. In: Lee, G., Howard, D., Ślęzak, D., Hong, Y.S. (eds.) ICHIT 2012. CCIS, vol. 310, pp. 699–706. Springer, Heidelberg (2012)
5. Kim, T.K., Lim, H.J., Chung, T.M.: Service Negotiation Model for response Time in Distributes Networks. Computing and Informatics, 395–405 (2004)
6. Grobauer, B., Walloschek, T., Stocker, E.: Understanding Cloud Computing Vulnerabilities. IEEE Security & Privacy 9(2), 50–57 (2011)
7. Kang, K.D., Son, S.H.: Dynamic Security and QoS Adaptation in Real-Time Embedded Systems. In: 26th IEEE Real-Time Systems Symposium, WIP Session, Miami, Florida (2005)
8. Kumar, K., Lu, Y.-H.: Yung-Hsiang Lu: Cloud Computing for Mobile Users: Can Offloading Computation Save Energy? Computer 43(4), 51–56 (2010)
9. Simoens, P., De Turck, F., Dhoedt, B., Demeester, P.: Remote Display Solutions for Mobile Cloud Computing. Computer 44(8), 46–53 (2011)
10. Zhang, P., Yan, Z.: A QoS-aware system for mobile cloud computing. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 15–17 (2011)

Intrusion-Tolerant Jini Service Architecture for Ensuring Survivability of U-Services Based on WSN

Sung-Ki Kim[1], Jae-Yeong Choi[3], Byung-Gyu Kim[2], and Byoung-Joon Min[3]

[1] Division of IT Education, Sun Moon University,
100 Kalsan-ri, Tangjeong-myeon, Asan-si, Republic of Korea 336-708
skkim@sunmoon.ac.kr
[2] Dep. of Computer Science and Engineering, Sun Moon University,
100 Kalsan-ri, Tangjeong-myeon, Asan-si, Republic of Korea 336-708
bg.kim@mpcl.sunmoon.ac.kr
[3] Dep. of Computer Science and Engineering, Incheon National University,
119 Academy-ro, Yeonsu-gu, Incheon, Republic of Korea 406-772
{jero,bjmin}@incheon.ac.kr

Abstract. U-Service environment based on WSN(Wireless Sensor Network) is poor in reliability of connection and has a high probability that the intrusion and the system failure may occur. Therefore, it is very important to ensure that the legitimate users make use of trustable services without discontinuance or obstacle of the services they are enjoying despite the presence of failures and intrusions. In this paper, we propose an intrusion-tolerant Jini service architecture integrating security and survivability mechanisms in order to provide end users with Jini services having a persistent state in wireless sensor networks. The proposed architecture is able to protect a Jini system not only from faults such as network partitioning or server crash, but also from attacks exploiting flaws. It is designed to provide performance enough to show a low response latency so as to support seamless service usage. Through the experiment on a test-bed, we have confirmed that the architecture is able to provide high security and availability at the level that the degradation of services quality is ignorable.

Keywords: Intrusion-tolerance, Jini, Apache River, Jgroup/ARM, Security.

1 Introduction

Jini[1], also called Apache River[2], is a java-based middleware supporting sharing of resources such as ubiquitous devices and software on networks while it copes with the heterogeneity of the lower levels such as the various types of devices or communication protocols. Jini provides a mechanism that discovers available services through the lookup services and makes a connection to the services that clients requested.

The networked systems based on WSNs(Wireless Sensor Networks) are apt to be partitioned due to a poor reliability of connection and have a high probability that the intrusion and the system failure may occur. Therefore, it is very important to guarantee the legitimate users make use of trustable services without discontinuance or obstacle of the services they are enjoying.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_19,

The Jgroup/ARM framework[3] has presented a middleware technology supporting the building of a dependable service in a distributed environment by introducing a concept is called Java-based object group platform. A set of distributed objects making a group take the responsibility for a service. It provides a good framework to solve the problem of the standard Jini systems that lack the supports for fault tolerant service developments. However, there are some shortcomings to apply this framework in the real environment directly. The first shortcoming is to lack the security provision for responding to the intrusion in the framework design. The system is made through this framework is greater danger than the standard Jini system because it is defenseless against intrusion due to no security mechanism and it can be easily compromised caused by group communication protocols performing share of service state data among replicated service objects as well. The second shortcoming is in which both the computation and communication cost are needed to merge the user service states among replicas are very high, when the partitioned replicas are merged into one group after a network is recovered.

In this paper, we propose an intrusion-tolerant Jini service architecture to support a seamless service usage while it establishes the dynamic trust in a Jini system and tolerates the system failures and intrusions. Section 2 briefly describes the contribution of the Jgroup/ARM systems and points out its shortcomings. Section 3 introduces our intrusion-tolerant Jini service architecture. Section 4 discusses the experimental results in terms of our contribution has introduced in section 3. Section 5 concludes the paper.

2 The Jgroup/ARM System

2.1 The Contribution of Jgroup/ARM System

In Fig. 1, from Fig. 1(a) to Fig. 2(d) illustrate service failure situations that may happen in the standard Jini service environment. In Fig. 1(a), all 65 users can make use of a Jini service without any problems. However, when a lookup server is crashed as shown in Fig. 1(b), all existing users keeping connections to the server can make use of the service within a given lease time. In the case of Fig. 1(c), 15 users lose all their connection and the opportunities that they can discover other services again. 30 users lose their connection to the server and only can rediscover unreachable service, and the presence of the service disappears from the lookup server. Only 20 users can make use of the service within a given lease time. In the case of Fig. 1(d), all users lose their connection to the server and only 30 users can rediscover unreachable service. In the case of Fig. 1(e), some users are subject to restriction on making use of services according to how a Jini network is partitioned but the situation that servers are crashed can overcome. In the case of Fig. 1(f), the Jini system overcomes failure events such as a specific server's crash and network partition as replicating the tasks providing the services on distributed computing nodes. The Jgroup/ARM system supports the case of Fig. 1(f). In Fig. 1(f), through considerable service resources are replicated, the Jgroup/ARM system can support the autonomous management of available resources so as to overcome all failure events through a dynamic service deployment and reconfiguration.

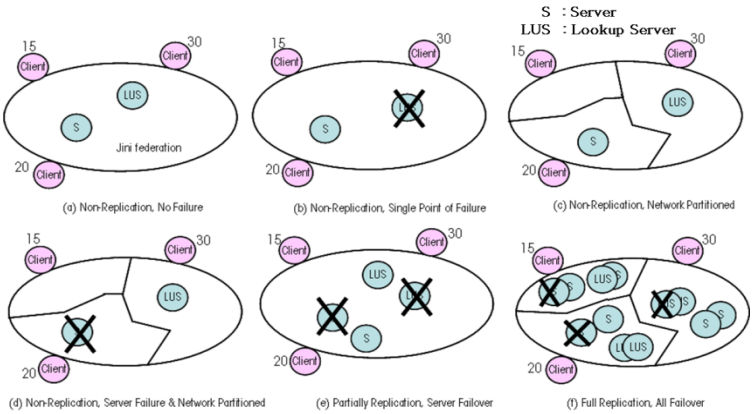


Fig. 1. Service Failure and Tolerance in a Jini Service Environment

2.2 A Definition and Examples for Discussion

Definition: In this section, we define the notion of a group of the distributed service objects for the further discussions as follows:

$$G_s \rightarrow \{R_1, R_2, R_3, \ldots R_n\}^e \quad (1)$$

Where,

- G_s : a server group that consists of a group of n replicated service objects
- R_n : the n-th replica which is a member of G_s
- e : the number of endpoints(i.e., service instances) that is created in each R_n in order to establish connection from clients under no failure condition that the system is not partitioned

Example 1: The expression, $G_s \rightarrow \{R_1, R_2, R_3 \ldots R_7\}^{20}$, means that all of 7 replicas takes the responsibility for a service and each replica has equally 20 connections from clients. This expression describes normal scenarios under no failure condition.

Example 2: The expression, $G_s \rightarrow \{\{R_1, R_2\}_a^3, \{R_3, R_4, R_5\}_b^7, \{R_6, R_7\}_c^{10}\}$, means that one server group is partitioned into 3 partitioned subgroups that take the responsibility for a service. This expression describes failure scenarios that the system is partitioned due to the intrusions and the system failures. In this scenario, replicas in each partitioned subgroup have endpoints enough to connect to their clients according to how a Jini network is partitioned. In this example, each replica in the subgroup ***a***, ***b***, and ***c***, has 3, 7 and 10 endpoints to connect to their client side proxy respectively.

2.3 The Shortcomings of Jgroup/ARM System

The Computational Overhead for Merging Distributed Service States of User: When the partitioned subgroups are merged into one group after the network failure is recovered, the Jgroup/ARM system faces a significant problem in bearing a heavy computation overhead for merging distributed service states of user. For recovery, the Jgroup/ARM system has a 2-phase merging operation as follows:

- Merging operation between leader replicas (for example, R_1-R_3-R_6)
- Merging operation between a leader and members (for example, R_1-R_2 , R_3-R_4-R_5 , R_6-R_7)
- The cost of this computational overhead can be estimated as follows:

$$\text{Time} - \text{Cost} = \sum_{i=1}^{p} \left\{ e_i^p * \left(e - e_i^p\right) * (mdt + c) \right\} \quad (2)$$

Where, $\boldsymbol{p}$ is the number of partitioned groups, $\boldsymbol{e_i^p}$ is the number of endpoints in each $\boldsymbol{R_n}$ of the $\boldsymbol{i}$-th partitioned subgroups after network is partitioned, and a $\boldsymbol{c}$ is a required time to make the service instances in each endpoint, and also ***mdt*** means multicast delay time for message delivery in group communication.

Latency in Callback Handler: In the Jgroup/ARM system, a group-proxy[4] for use of a service is offered, a selected proxy in the group of proxies sends the requests to server-side replica corresponding to it. The selected replica replies their service result on behalf of all of members after the callback handler in the selected replica receives the service results from all members. This latency in callback handler operation is a cost to pay for fault detection and failover provision.

The Lack of the Security Mechanism: The group communication among replicas relies on the Jgroup daemon(JD) in every node hosting services[4]. However, the JD supporting group multicast communication does not ensure the confidentiality and integrity of messages. In addition, there is no authentication mechanism providing the trust among replicas. Thus, attackers can easily intrude into the system by tempering all communication messages and configuration files.

3 Our Proposed Architecture

3.1 An Extension to Session-Based Allocation of Replica Instances

In comparison with the Jgroup/ARM architecture, one of extended features in our proposed architecture is to support the identification and management of endpoints in each replica on the basis of the session while maintaining group membership among replicas regardless of merging subgroups. In our system, a smart proxy in the client-side manages the user service states whenever it receives response messages in order to solve the problem above. Fig. 1 and Fig. 2 show our architecture and contribution in comparison with Jgroup/ARM system through example 2 and expression (2) respectively.

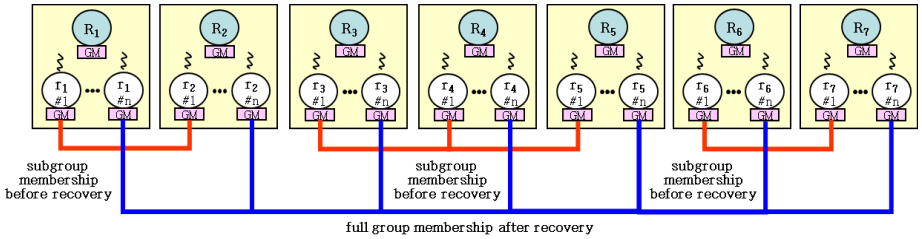


Fig. 2. Session-based Allocation of Replica Instances and Their Membership Management

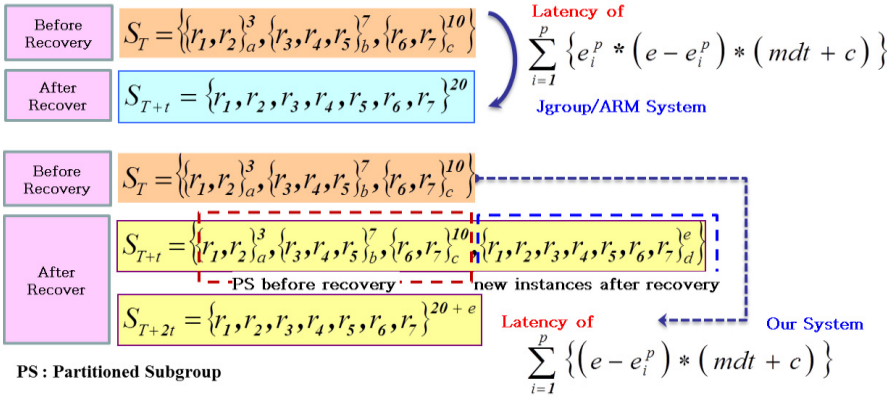


Fig. 3. Reduction of Time-cost for Merging the User Service States

Table 1 illustrates Fig. 2 above. In Table 1, *U* means users who send requests through client for use of a service, and the prefix ***i*** indicates a number to identify individuals. The cost of computational overhead of our system can be estimated as follows:

$$\text{Time} - \text{Cost} = \sum_{i=1}^{p}\{(e - e_i^p) * (mdt + c)\} \quad (3)$$

Table 1. An Simple Simulation to Illustrate Fig.3 above

Replica	No Failure Condition	Before Recovery	After Recovery	Absentee Data	Leadercast Contributor
R_1 R_2	U_1 U_1	U_1 , U_2 U_1 , U_2	U_1 , U_2, U_5, U_1 , U_2, U_5,	U_3 , U_4	R_3 , R_6
R_3 R_4 R_5	U_1 U_1 U_1	U_1 , U_3 U_1 , U_3 U_1 , U_3	U_1 , U_3, U_5, U_1 , U_3, U_5, U_1 , U_3, U_5,	U_2 , U_4	R_1 , R_6
R_6 R_7	U_1 U_1	U_1 , U_4 U_1 , U_4	U_1 , U_4, U_5, U_1 , U_4, U_5,	U_2 , U_3	R_1 , R_3

3.2 Callback Handler for Intrusion Tolerance

Our proposed architecture applies the byzantine agreement algorithm [5-6] and design diversity to service implementation in order to mask results introduced from compromised server due to intrusions. To accomplish this goal, we have added an additional function to callback handler in order to act as a voter masking the compromised results. When the number of total replica is N and the number of compromised replica is T of N, if there are replicas of more than 2/3 N to satisfy a condition that $N > 3T$, callback handler replies the service results to client.

3.3 Security Architecture for Secure Communication

Our proposed architecture supports security provisions to satisfy the goals for providing secure Jini services such as authentication, access control, and confidentiality as follows:

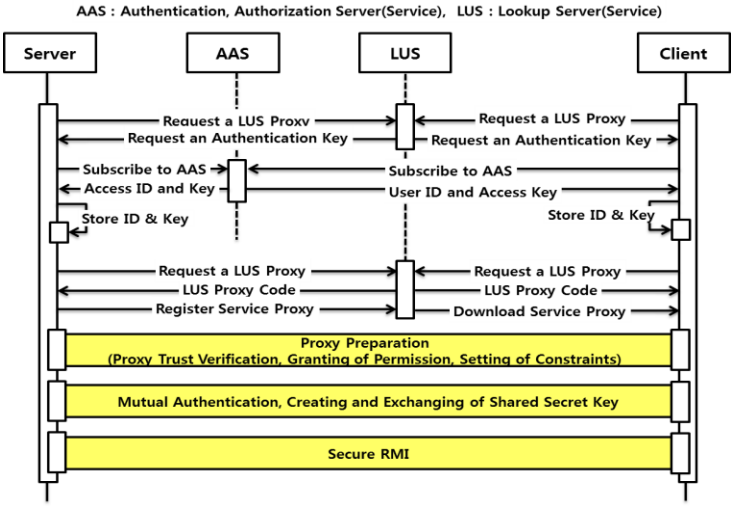


Fig. 4. Architecture for Secure Communication among Service Entities

4 The Discussion on Experimental Results

4.1 Performance Analysis

We have experimented with 2 measures of state merging times and turnaround times. Through a several measurement, It has been shown that the time cost of our system is 2 times lower than one of Jgroup/ARM system in proportional to e of expression(1), when a recovery operation to merge the user service states has been performed under system-partitioned condition(3 replica per partitioned group in the left in Fig.5). The right in Fig. 5 shows measurement results of turnaround time in concern with

leadercast, multicast and voting applications. According to the manner of delivering messages, a request message from a client is delivered to the replicated services and then a secure service result from them is committed through a voting process.

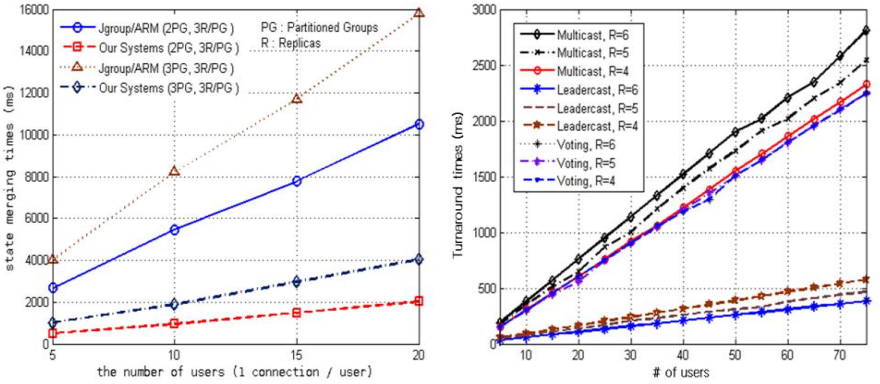
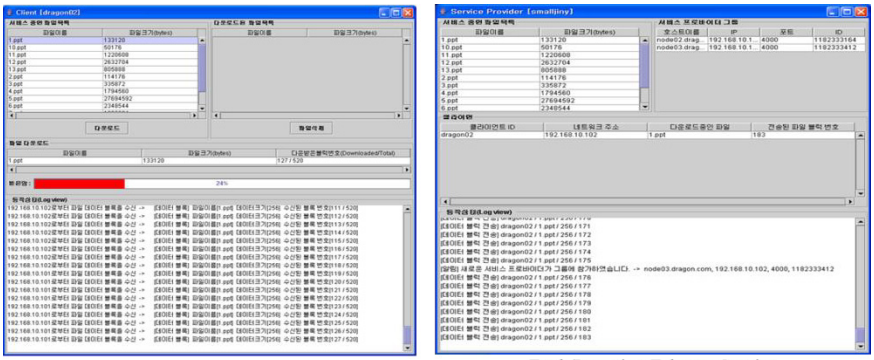


Fig. 5. Time Cost Merging User Service States(left) and Turnaround Time Cost(right)

4.2 The Implementation Results

Fig.6 shows a snapshot that a file transfer service is keeping up its mission in spite of a server-crashed failure. In this experiment, when a system consists of one client and two replicated services in 100Mbps Ethernet environment, it has been shown that the failover latency is 340 ms(160 ms for failure detection + 180 ms for receiving stateful service from other replica) for seamless service usage.



Normal Operation under No Failure Condition

Fault/Intrusion Tolerant Service after Server-Crashed Failure

Fig. 6. A Snapshot to Check the Seamless Service Usage

5 Conclusion

U-Service environment based on WSN is poor in reliability of connection and has a high probability that the intrusion and the system failure may occur. In this paper, we propose an intrusion-tolerant Jini service architecture integrating security and survivability mechanisms in order to provide end users with trustable Jini services having a persistent state in WSNs. The proposed architecture is able to protect a Jini system not only from faults such as network partitioning or server crash, but also from attacks exploiting flaws. It is designed to provide performance enough to show a low response latency so as to support seamless service usage. We believe that our proposed architecture is a good reference model for building a better secure U-service infrastructure.

References

1. Apache Software Foundation, Apache Jini Specifications v2.1.2, Published Specification, http://river.apache.org/doc/spec-index.html
2. Apache Software Foundation, Apache River User Guide, http://river.apache.org/user-guide-basic-river-services.html
3. Meling, H., et al.: Jgroup/ARM: a distributed object group platform with autonomous replication managements. Software Practice and Experience (2007)
4. Kolltveit, H., Hvasshovd, S.-O.: Preventing Orphan Requests by Integrating Replication and Transactions. In: Ioannidis, Y., Novikov, B., Rachev, B. (eds.) ADBIS 2007. LNCS, vol. 4690, pp. 41–54. Springer, Heidelberg (2007)
5. Pease, M., Shostak, R., Lamport, L.: Reaching Agreement in the Presence of Faults. Journal of the ACM 27(2), 228–234 (1980)
6. Min, B.J., Kim, S.K., Im, C.: Committing Secure Results with Replicated Servers. In: Laganá, A., Gavrilova, M.L., Kumar, V., Mun, Y., Tan, C.J.K., Gervasi, O. (eds.) ICCSA 2004. LNCS, vol. 3043, pp. 246–253. Springer, Heidelberg (2004)

Creation Mechanism for Access Group Based on User Privacy Policy-Based Protection*

Taekyeong Kang[1], Hyungkyu Lee[2], Dong-Hwan Park[2], Hyo-Chan Bang[2], and Namje Park[1,**]

[1] Major in Elementary Computer Education, Department of Primary Education, Graduate School of Education, Jeju National University, 61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Korea
{ktg,namjepark}@jejunu.ac.kr
[2] Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, Korea
{leehk,dhpark,bangs}@etri.re.kr

Abstract. Smart grid is the next-generation intelligent power network which optimizes energy efficiency through the mutual real time exchange of information between power supplier and consumer through the integration of existing power network and the information technology (IT). However, the smart grid environment can have problems involved with personal privacy invasion. This paper suggests the creation mechanism of privacy policy-based protection system.

Keywords: Smart Grid Security, Privacy Policy, Privacy Protection, Access control, Privacy Exchange Format.

1 Introduction

Smart grid is a new electricity grid which transmits and distributes electricity intelligently by converging the information technology into the traditional electricity grid. Recently, the smart grid projects are promoted rapidly because the 'Green IT' becomes more and more interesting. However, Modernization of the grid will increase the level of personal information detail available as well as the instances of collection, use and disclosure of personal information.

In this paper, we propose privacy policy-based protection system based on smart grid environment. The structure of the privacy policy-based protection system using load management system in the smart grid environment is the structure that serves data in the load management system to the web through the application service network. For this, the privacy policy-based protection system suggested and

* This work was supported by the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy (MKE, Korea) [10038653, Development of Semantic Open USN Service Platform].

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_20,

developed the smart grid privacy policy-based protection system which controls service access by protecting items related to the personal information of the user and setting the privacy protection level for each item. Also, it analyzes the outcomes of smart grid privacy policy-based protection system. By applying smart grid privacy policy-based protection system, the user can show his/her information to the users requesting for such information through the mobile device or PC based on the privacy level he/she personally set and the information related with the device or to the users he/she designate and also receive the information he/she wants when he/she wants the information. By using this system, the company providing application services will be able to protect the personal information and become a reliable company and provide the service the user wants based on the information collected, expecting greater sales.

2 Overview of Policy-Based Protection

The privacy policy-based protection system is a service where the user who owns the information provides the mechanism for protecting his/her privacy. It is composed of the privacy policy-based protection system which manages the user's privacy protection policy, the system which determines the privacy policy of the user and sends this to the privacy policy-based protection system, and the system which provides information based on user's privacy protection policy.

3 Initial Privacy Creation Mechanism

Initial privacy is created based on the initial privacy policy which defines the privacy policy, the item that exists in each service, by combining initial privacy policy and the administrator's information application schema. The example that uses this mechanism is presented in Figure 1.

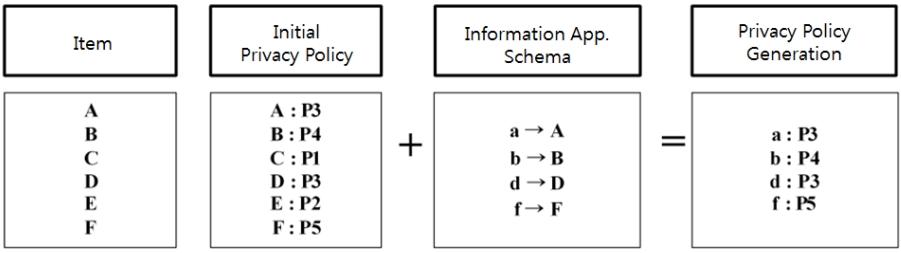


Fig. 1. Initial privacy generation mechanism

The item lists the type of information handled by each service group and the initial privacy policy provides basic privacy level recommended by the certificate authorities depending on information type, and this value is represented by the combination of the letter "P" and the number. Higher number indicates that the privacy must be protected more intensively. The information application schema is the list of the types

of information that the administrator handles among the information handled by the item, and the administrator must map the information he/she handles with the information listed in the item. The type of information handled in this example is labeled with lower case English letter, and is mapped one-to-one with the information in the item to form the information application schema. A unique initial privacy can be created by using information application schema and initial privacy policy.

4 User Privacy Creation Mechanism

Like initial privacy, user privacy is created based on the initial privacy policy. First, the list of information handled enumerated by initial privacy policy and the default privacy level for information is suggested to the user. The user can check the default value and change each value if necessary or keep the default value. User privacy policy is the list that reflects user's intention determined as above. The user privacy is created by mapping this user privacy policy and the information application schema of each user.

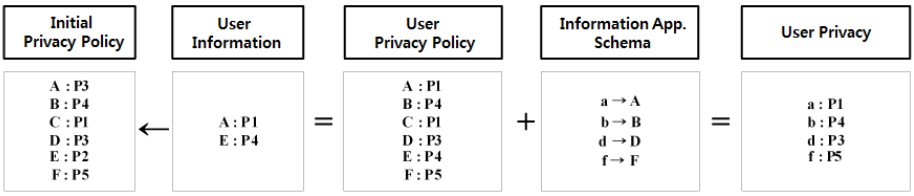


Fig. 2. User privacy generation mechanism

The example in Figure 2 shows a process where a user creates his/her privacy policy for the information belonging to a service group. The privacy policy-based protection system allows the user to make decisions by services handling similar information while automatically creating and distributing the user privacy of the user to be given to this information whenever there is user input.

In the example, the user changed the privacy policy for A and E concerning the service group and the user privacy policy of the user is created by reflecting this value. Also, the user privacy relating to the mapping with the user privacy policy and information application server schema is created and transmitted since the information application schema is different.

5 Access Group Creation Mechanism

The purpose of the access group is to allow the owner of information to authorize the person he/she selects to access his/her information by designating the mobile phone number of the person to authorize to the designated access authority level. Currently, the privacy policy-based protection system allows the user to set access authority

level from 0 ~ 9, and larger number indicates higher access authority for information. When the privacy policy-based protection system receives user input, it creates a security token based on the input, and currently SHA1 is used as the algorithm for creating a security token. The access group is divided into the user access group which provides the security token created based on user input, the initial access group which provides the token to be used as the default value for each access group when there is no user defined value, and the privacy level which designates the level of the information to be given as default value to those who are not designated by the user. 1 access group of 1 user is given to each service group, and it is newly created and sent to the companies included in each service each time the user defined value changes. For the request for the access to user information, each company controls the access through the mechanism of comparing the security token provided with the request and the security token presented to the access group.

6 Conclusion

The invasion of privacy in the smart grid involves high risk of disclosing personal information from the level of collecting trivial information up to the level of the individual behavioral pattern based on the information collected. For this, the invasion of privacy is becoming the center of interest and a number of studies have been made to resolve this issue.

In this paper, the privacy policy-based protection system is designed to allow the user to directly set the privacy level of each item and designate the access level for the third party to block the access of undesignated user and protect important information.

References

1. Lee, J.W., Park, N.: Individual Information Protection in Smart Grid. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 153–159. Springer, Heidelberg (2012)
2. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
3. Park, N.: Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
4. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
5. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)

6. Park, N., Kim, Y.: Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS, vol. 6422, pp. 193–202. Springer, Heidelberg (2010)
7. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS, vol. 6422, pp. 142–152. Springer, Heidelberg (2010)
8. Park, N.: Customized Healthcare Infrastructure Using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
9. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods for Knowledge Management and Communication. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
10. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
11. Park, N.: The Implementation of Open Embedded S/W Platform for Secure Mobile RFID Reader. The Journal of Korea Information and Communications Society 35(5), 785–793 (2010)
12. Park, N., Cho, S., Kim, B.-D., Lee, B., Won, D.: Security Enhancement of User Authentication Scheme Using IVEF in Vessel Traffic Service System. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 699–705. Springer, Heidelberg (2012)
13. Ko, Y., An, J., Park, N.: Development of Computer, Math, Art Convergence Education Lesson Plans Based on Smart Grid Technology. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 109–114. Springer, Heidelberg (2012)
14. Ko, Y., Park, N.: Experiment and Verification of Teaching Fractal Geometry Concepts Using a Logo-Based Framework for Elementary School Children. In: Kim, T.-H., Adeli, H., Slezak, D., Sandnes, F.E., Song, X. (eds.) FGIT 2011. LNCS, vol. 7105, pp. 257–267. Springer, Heidelberg (2011)
15. An, J., Park, N.: The Effect of EPL Programming Based on CPS Model for Enhancing Elementary School Students' Creativity. In: Park, J.J. (J.H.), Jeong, Y.-S., Park, S.O., Chen, H.-C. (eds.) EMC Technology and Service. LNEE, vol. 181, pp. 237–244. Springer, Heidelberg (2012)
16. An, J., Park, N.: Computer Application in Elementary Education Bases on Fractal Geometry Theory Using LOGO Programming. In: Park, J.J., Arabnia, H., Chang, H. (eds.) ITCS & IRoA 2011. LNEE, vol. 107, pp. 241–249. Springer, Heidelberg (2011)
17. Park, N., Ko, Y.: Computer Education's Teaching-Learning Methods Using Educational Programming Language Based on STEAM Education. In: Park, J.J., Zomaya, A., Yeo, S.-S., Sahni, S. (eds.) NPC 2012. LNCS, vol. 7513, pp. 320–327. Springer, Heidelberg (2012)
18. Kim, Y., Park, N.: Development and Application of STEAM Teaching Model Based on the Rube Goldberg's Invention. In: Yeo, S., Pan, Y., Lee, Y.S. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 693–698. Springer, Heidelberg (2012)
19. Kim, Y., Park, N.: The Effect of STEAM Education on Elementary School Student's Creativity Improvement. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech, CA, CES3 2012. CCIS, vol. 339, pp. 115–121. Springer, Heidelberg (2012)

20. Kim, Y., Park, N.: Elementary Education of Creativity Improvement Using Rube Goldberg's Invention. In: Park, J.J. (J.H.), Kim, J., Zou, D., Lee, Y.S. (eds.) ITCS & STA 2012. LNEE, vol. 180, pp. 257–263. Springer, Heidelberg (2012)
21. Kim, K., Kim, B.-D., Lee, B., Park, N.: Design and Implementation of IVEF Protocol Using Wireless Communication on Android Mobile Platform. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 94–100. Springer, Heidelberg (2012)
22. Kim, G., Park, N.: Program Development of Science and Culture Education Tapping into Jeju's Special Characteristics for Adults. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech, CA, CES3 2012. CCIS, vol. 339, pp. 133–138. Springer, Heidelberg (2012)
23. Hong, J., Park, N.: Teaching-learning Methodology of STS Based on Computer and CAI in Information Science Education. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 733–738. Springer, Heidelberg (2012)
24. Kim, M., Park, N., Won, D.: Security Improvement on a Dynamic ID-Based Remote User Authentication Scheme with Session Key Agreement for Multi-server Environment. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech, CA, CES3 2012. CCIS, vol. 339, pp. 122–127. Springer, Heidelberg (2012)
25. Kim, Y., Park, N.: Customizing Data Analysis Using Forensic Approach in Personalized RFID for Access Control of Student's Harmful Contents. In: Park, J.J. (J.H.), Kim, J., Zou, D., Lee, Y.S. (eds.) ITCS & STA 2012. LNEE, vol. 180, pp. 249–256. Springer, Heidelberg (2012)
26. Lee, B., Park, N.: Security Architecture of Inter VTS Exchange Format Protocol for Secure u-Navigation. In: Park, J.J. (J.H.), Jeong, Y.-S., Park, S.O., Chen, H.-C. (eds.) EMC Technology and Service. LNEE, vol. 181, pp. 229–236. Springer, Heidelberg (2012)
27. Lee, J., Bang, H., Park, N.: Design of Mobile NFC Extension Protocol for Various Ubiquitous Sensor Network Environments. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 715–722. Springer, Heidelberg (2012)
28. Lee, J.W., Kim, G., Park, N.: A Study on the Learner Participation Interaction for Multi-modality Integration for Smart Devices. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech, CA, CES3 2012. CCIS, vol. 339, pp. 139–143. Springer, Heidelberg (2012)
29. Lee, J., Park, N.: Encryption Scheme Supporting Range Queries on Encrypted Privacy Databases in Big Data Service Era. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 739–746. Springer, Heidelberg (2012)

Specification of Train Control Systems Using Formal Methods

Bingqing Xu[1] and Lichen Zhang[2,*]

[1] Software Engineering Institute, East China Normal University, 200062 Shanghai, China
xbqjoya@gmail.com
[2] Faculty of Software Engineering Institute, East China Normal University,
Shanghai 200062, Shanghai, China
zhanglichen1962@163.com

Abstract. Just as what the public pursue, we need a much safer railway system with a higher level of automation in control. To achieve this goal, the author aims to specify the Train Control System by formal methods which can specify the communication of various processes in the system clearly. This paper applies Timed-CSP which concerns the time-delay to the specification of the control flow and communication among flows in Train Control System, and specifies the state and data change by Object-Z. By Timed-CSP and Object-Z, the specification of a simplified Train Control System especially the time constraints is more concrete.

Keywords: Timed-CSP, Object-Z, specification, control and sensor.

1 Introduction

Complicated system such as train control system is a system with many complex behavioural aspects. With help of formal methods, we now find a way to construct a detailed specification of each aspect and the link mechanism among various aspects. While a communication mechanism is not enough to describe both the state change and data change in the system. Concerning the time characteristics in the system, Timed-CSP contains new notations which can specify the system with time constraints better. In addition, Object-Z is ideal for analysis in data changes.

1.1 CSP

Communicating Sequential Processes (CSP) is a formal language for describing patterns of interaction in concurrent systems. And it is a mathematical theories which is known as process algebras. In 1978, it is first described by C. A. R. Hoare. CSP has been applied in industry as a tool for specifying and verifying the concurrent aspects of a variety of different systems. Its syntax is as below, detailed explanation is in [1-2].

$$P :: STOP|SKIP|a \rightarrow P|P \square P|P \sqcap P|\ P; P \mid P \parallel P|\ P||\ |\ P \mid f(P) \mid f^{-1}(P) \mid P\backslash A \mid \mu X \cdot F(X)$$

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_21, © Springer-Verlag Berlin Heidelberg 2014

1.2 Timed-CSP

Timed-CSP is an extension of Communicating Sequential Processes (CSP). It was first described in 1986 by Reed and Roscoe and its main difference between CSP is that it proposes the idea **WAIT t** for time-delay. Later in this paper, **WAIT t** is applied in prefixing and sequential composition. Its syntax is as below, detailed explanation is in [2-3].

$$P ::= \perp \mid STOP \mid SKIP \mid WAIT\ t \mid a \rightarrow P \mid P \quad P \mid P \sqcap P \mid P; P \mid P \parallel P \mid P \mid\mid\mid P \mid$$
$$P_A \parallel_B P \mid f(P) \mid f^{-1}(P) \mid P \backslash A \mid \mu X \cdot F(X)$$

1.3 Object-Z

The language Z is a mathematical notation with a schema form which is developed by Abrial and others at the Programming Research Group of the Oxford University Computing Laboratory (OUCL) since the late 1970s [4]. Since the influence of object-oriented paradigm, Obeject-Z (OZ) emerges as the extension of language Z. In OZ, the schema box is a class, all the variables are grouped in the box and operation can be also described in the schema with pre and post condition. Details of OZ are in [5].

2 Case Study: Train Control System

To illustrate this control system by Timed-CSP and Object-Z, the author presents the case study. The following Figure 1 clearly indicates the track elements of the Train Control System which includes crossing area, linear tracks, controllers and sensors. In the Train Control System, we have ***n*** trains on the track synchronously, and trains have various ***id***s, directions and destinations. The sensors collect the real-time information of the trains and keep them for the controller to send signals to the train. When trains are near the crossing area, or have any type of emergency, sensors can detect it and keep information in database for a better real-time control. To achieve a high level of automation, trains can only receive, report and feedback information.

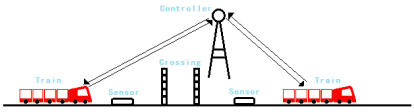


Fig. 1. Train control system

In Figure 2, it is obvious to see that the whole system is composed of three main aspects in function: ***TrainController***, ***Sensor***, and ***Train***. Each aspect consists of several sub-functional modules.

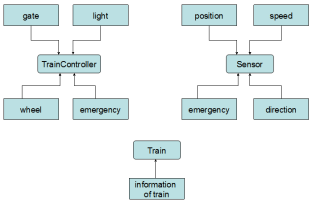


Fig. 2. Functional modules in train control system

In Figure 3, it shows the communication methods between any two aspects. The name of operation is above the arc.

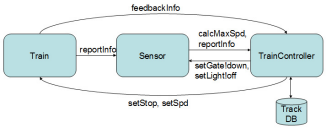


Fig. 3. Communication among the train control system modules

The author specifies the case in two sections: (i) Specification of Train Control System by Timed-CSP; (ii) State and data changes by Object-Z and combination with Timed-CSP.

Timed-CSP

Train All the trains from $\mathbf{Train_1}$ to $\mathbf{Train_n}$ have the same alphabet as follows. **t1** denotes the time-delay between action in the same processes, while **t2** denotes the time-delay between different processes.

$$Train_1 \overset{c}{=} reportInfo \overset{t2}{\rightarrow} Sensor|feedbackInfo \overset{t2}{\rightarrow} TrainController|receiveInfo \overset{t1}{\rightarrow} Train$$

$$\cdots\cdots$$

$$Train_n \overset{c}{=} reportInfo \overset{t2}{\rightarrow} \text{Sensor}|feedbackInfo \overset{t2}{\rightarrow} TrainController|receiveInfo \overset{t1}{\rightarrow} Train$$

Sensor Process ***Sensor*** consists of ***NormalSensor*** and ***EmergSensor*** which indicate normal event sensor and emergency sensor. ***NormalSensor*** gets speed, position and direction of the train, and then calculate the maximum speed. The two sub-processes interleave from time to time. **t1** denotes the time-delay between action in the same processes, while **t2** denotes the time-delay between different processes.

$$\text{Sensor} \overset{c}{=} NormalSensor ||| \text{ EmergSensor}$$

$$\text{NormalSensor} \overset{c}{=} receiveInfo \overset{t1}{\rightarrow} NormalSensor \left| reportInfo \overset{t2}{\rightarrow} TrainController \right| \text{getSpd} \overset{t1}{\rightarrow} getPos \overset{t1}{\rightarrow} getDir \overset{t1}{\rightarrow} calcMaxSpd \overset{t2}{\rightarrow} TrainController$$

$$\text{EmergSensor} \overset{c}{=} detectEmergency \overset{t1}{\rightarrow} EmergSensor$$

TrainController Process ***TrainController*** consist of ***NormalControl*** and ***EmergControl*** which indicate normal event control and emergency control. ***NormalControl*** indicates speed, wheel, light, gate control whenever the train is on the linear track or in the crossing area. When the crossing area is safe and a train wants to enter, it is allowed to go and ***NormalControl*** sets the light yellow then red, and the gate is set down to prevent the other trains. But when the train leaves the area, gate is set up and the light is off. ***EmergControl*** sends signal to train when it receives emergency sensor. The two sub-processes interleave from time to time. **t1** denotes the time-delay between action in the same processes, while **t2** denotes the time-delay between different processes.

$$\text{TrainController} \overset{C}{=} \text{NormalControl} ||| \text{ EmergControl}$$

$$\text{NormalControl} \overset{C}{=} setSpd \overset{t2}{\rightarrow} Train | (setLight!\, yellow \overset{t1}{\rightarrow} setLight!\, red \overset{t1}{\rightarrow} setGat!\, down | setGat!\, up \overset{t1}{\rightarrow} setLight!\, off) \overset{t2}{\rightarrow} Sensor$$

$$\text{EmergControl} \overset{C}{=} \text{receiveInfo} \overset{t1}{\rightarrow} setStop \overset{t2}{\rightarrow} Train$$

As the above Figure 3 shows, the three processes interact, and **WAIT t2** denotes the time-delay between different processes.

$$\text{Train; TrainContoller} \overset{C}{=} Train; WAIT\ t2; TrainController$$

$$\text{Train; Sensor} \overset{C}{=} Train; WAIT\ t2; Sensor$$

$$\text{Sensor; TrainController} \overset{C}{=} Sensor; WAIT\ t2; TrainController$$

$$\text{TrainController; Sensor} \overset{C}{=} TrainController; WAIT\ t2; Sensor$$

$$\text{TrainController; Train} \overset{C}{=} TrainController; WAIT\ t2; Train$$

Totally, ***MAIN*** process is as follows:

$$\text{Main} \overset{C}{=} (Train_1 \parallel \cdots \parallel Train_n) \parallel TrainController \parallel Sensor \parallel OtherParallelProcesses$$

Object-Z

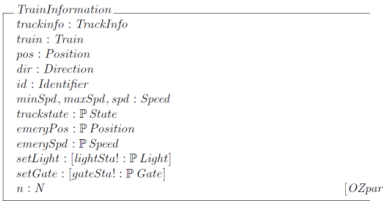
In the schema box, paper generates the CSP part and Object-Z part together to form a new definition. Firstly, CSP part is up above.

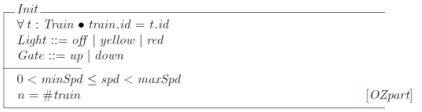
MainTrainController

$Main \overset{c}{=} (Train_1 \parallel \cdots \parallel Train_n) \parallel TrainController \parallel Sensor \parallel OtherParallelProcesses$
$Train_1 \overset{c}{=} reportInfo \xrightarrow{t2} Sensor \mid feedbackInfo \xrightarrow{t2} TrainController \mid receiveInfo \xrightarrow{t1} Train$
$\vdots$
$Train_n \overset{c}{=} \cdots$

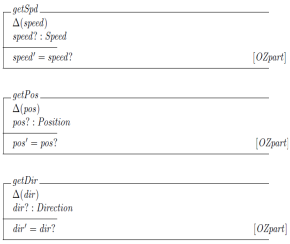
$Sensor \overset{c}{=} NormalSensor \,|||\, EmerySensor$
$NormalSensor \overset{c}{=} \cdots \qquad EmerySensor \overset{c}{=} \cdots$
$TrainController \overset{c}{=} NormalControl \,|||\, EmeryControl$
$NormalControl \overset{c}{=} \cdots \qquad EmeryControl \overset{c}{=} \cdots$
$Train;\ TrainController \overset{c}{=} Train;\ WAIT\ t2;\ TrainController$
$Train;\ Sensor \overset{c}{=} \cdots \qquad Sensor;\ TrainController \overset{c}{=} \cdots$
$TrainController;\ Sensor \overset{c}{=} \cdots \qquad TrainController;\ Train \overset{c}{=} \cdots$ [CSPpart]

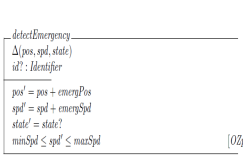
For each train, there is a series of information. For instance: ***id***, ***pos***, ***dir***, ***spd*** etc. And there are many operations such as ***setLight*** and ***setGate***. And the *Object-Z (OZ)* part begins at ***Init*** class that initiates the parameters.

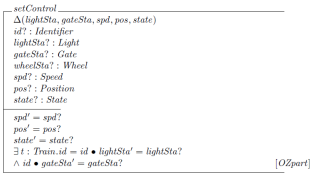




The following class ***getSpd***, ***getPos***, ***getDir***, ***detectEmergency***, ***setControl*** are classes which define the operations. Notation Δ means that the parameter after it is the parameter will change its value, other parameters will not change. And author gives the precondition and post condition of the parameter value that may change. "**?**" indicates the input value and "**'**" indicates the value after change.







3 Related Work

J Faber et.al gives a general description of the radio block center which communicates with all the train on the track with CSP-OZ-DC [6]. J Hoenicke divides the controller into several parts, and aims to guarantee an emergency control and remote control of points and crossing area in the railway system with CSP-OZ-DC [7-8]. And J Faber et.al simply analyzes the train on the linear track with CSP-OZ-DC [9]. Since many researchers have talked about the train control system, though those adopt similar schema box to illustrate the specification, time characteristic is not clearly pointed out such as papers [6-7].

4 Conclusion

This paper applies Timed-CSP and Object-Z to the specification of the Train Control System. Different from other related work, this paper concerns the time-delay not only in the single process, but also time-delay between various processes. And state specification by Object-Z is a plus to the whole control system which defines the data changes. The combination of these two formal methods in a schema box clearly separates the CSP part and OZ part, which effectively avoids misunderstanding as well. But paper wishes to give a more precise description of physical part of the train, such as the method to control speed in further study in the future.

Acknowledgments. This work is supported by Shanghai Knowledge Service Platform Project (No.ZF1213), national high technology research and development program of China (No.2011AA010101), national basic research program of China (No.2011CB302904), the national science foundation of China under grant (No. 61173046, No.61021004, No.61061130541, No.91118008), doctoral program foundation of institutions of higher education of China (No.20120076130003), national science foundation of Guangdong province under grant (No.S20110100 04905).

References

1. Davies, J.: Specification and Proof in Real-Time Systems, Oxford, England (1993) ISBN 0-902928-71-6
2. Roscoe, A.W.: The Theory and Practice of Concurrency. Prentice-Hall, Pearson (2005)
3. Ouaknine, J.: Timed CSP: A Retrospective. Electronic Notes in Theoretical Computer Science 162, 273–276 (2006)
4. Hoenicke, J.: Combination of Process, Data, and Time, Oldenburg, Germany (2006) ISSN 0946-2910
5. Sühl, C.: An Integration of Z and Timed CSP for Specifying Real-Time Embedded Systems (2002)
6. Faber, J., Jacobs, S., Sofronie-Stokkermans, V.: Verifying CSP-OZ-DC Specifications with Complex Data Types and Timing Parameters. In: Davies, J., Gibbons, J. (eds.) IFM 2007. LNCS, vol. 4591, pp. 233–252. Springer, Heidelberg (2007)
7. Hoenicke, J.: Specification of Radio Based Railway Crossings with the Combination of CSP, OZ, and DC. LNCS (2001)
8. Hoenicke, J., Olderog, E.-R.: Combining specification techniques for processes, data and time. In: Butler, M., Petre, L., Sere, K. (eds.) IFM 2002. LNCS, vol. 2335, pp. 245–266. Springer, Heidelberg (2002)
9. Faber, J., Ihlemann, C., Jacobs, S., Sofronie-Stokkermans, V.: Automatic Verification of Parametric Specifications with Complex Topologies. In: Méry, D., Merz, S. (eds.) IFM 2010. LNCS, vol. 6396, pp. 152–167. Springer, Heidelberg (2010)

Formal Descriptions of Cyber Physical Systems Using Clock Theory

Bingqing Xu[1] and Lichen Zhang[2,*]

[1] Software Engineering Institute, East China Normal University, 200062 Shanghai, China
xbqjoya@gmail.com
[2] Faculty of Software Engineering Institute, East China Normal University,
Shanghai 200062, Shanghai, China
zhanglichen1962@163.com

Abstract. Cyber Physical Systems are in charge of the control of physical processes characterized by their own dynamics. This control must comply with timing constraints - sometimes stringent ones- imposed by the Cyber Physical Systems. It is crucial to address these timing issues as early as possible in the development process to detect inconsistencies in the requirements or in the constraints and to capture changes in the system. This paper aims to apply the clock theory to the specification of Cyber Physical Systems. To illustrate the concept we develop a well-known case study: the Steam Boiler Control System.

Keywords: Cyber Physical Systems, continuous·discrete, clock theory,·time analysis.

1 Introduction

In Cyber Physical systems [1-3], time constraints [4] are vital elements which influence the description of physical process and control procedure. Cyber Physical Systems are newly defined dynamic systems which exhibit both continuous and discrete dynamic behaviour. For all those discrete control events and continuous time flows, the bilateral interaction between them remains. Therefore, time constraint becomes a crucial part for a formal description as it describes the mechanism of the link rule between discrete events and continuous time flows. This paper aims to consider the timing issues in Cyber Physical Systems and give a formal specification of time in Cyber Physical Systems by a case study. In order to guarantee the consistency of time constraints, this paper would prefer to apply the unified clock which is put forward in clock theory to the specification of the cyber physical systems.

The already existing methods for specification of system are various and each has its own merits, but the specification of interaction among subsystems is not so complete. Clock theory[5] has a detailed specification on time, description of dynamic

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_22,

operations of continuous world, and each event is combined with a clock which describes its time constraint. Leeb and many other researchers have done lots of work on the specification of steam boiler control system [6], and clock theory is ideal for specification of the system which requires continuous steam boiler and discrete controller as well.

2 Clock Theory

Clock theory [5] puts forward the possibility to record, describe, and analyze events in physical world with a clock. Thus time description is clearer in every event and the theory sets up a better link between continuous world and discrete world.

Definition 1. A clock ***c*** is an increasing sequence of real numbers, ***c[1]*** stands for the first element of ***c***. We define its low and high rates by:

$$\Delta(c) =_{df} \boldsymbol{inf}\{(c[i+1]-c[i]) \mid i \in Nat\} \tag{1}$$

$$\nabla(\mathrm{c}) =_{\mathrm{df}} \sup\{(\mathrm{c[i+1]}-\mathrm{c[i]}) \mid \mathrm{i} \in \mathrm{Nat}\} \tag{2}$$

Definition 2 (Partial Order in Clock). If ***c*** runs faster than ***d***. For all $\mathrm{i} \in$ Nat, $\mathrm{c[i]} \leq \mathrm{d[i]}$. Then relation of ***c*** and ***d*** can be denoted by:

$$\mathrm{c} \preccurlyeq \mathrm{d} \tag{3}$$

Definition 3. Let ***c*** and ***d*** be clocks. We define the transition latency between the two clocks as

$$\rho(\mathrm{c},\mathrm{d}) =_{df} \sup\{|c[i]-d[i]| \mid i \in Nat\} \tag{4}$$

Definition 4. Some dynamic features of continuous variable can be described better as follows. ***climb(u,r)*** is introduced to describe the time instants when the value of u rises up to r. ***drop(u,r)*** is introduced to describe the time instants when the value of u falls below r.

Definition 5. (Linking Mechanism). Here ***c*** is a clock with c[1]>0, and $\boldsymbol{X_0}$ is an initial value. We assign the value of continuous variable ***u*** to discrete variable ***x*** at the every instant of clock ***c***.

$$\mathrm{x} = \mathrm{u} \text{ every c init } x_0 \tag{5}$$

In differential equation, ***u*** is a continuous variable, $\boldsymbol{u_0}$ is an initial value, and ***f*** is an expression. $(\dot{\mathrm{u}} = \mathrm{f}) \wedge (\mathrm{u}(0) = \mathrm{u}_0)$ is the relation between ***u*** and ***f***, then

$$\dot{\mathrm{u}} = f \; init \; u_0 \tag{6}$$

Let ***e*** be an event. **clock**(e) denotes the clock that records the time instants when ***e*** occurs. And **clock**(**event**(c)) denotes the time instants that the event takes place at every time instant c[i].

3 Case Study: Steam Boiler Control System

As Figure 1 shows, the steam boiler control system [6] consists of two parts. One is the steam boiler which acts continuously, and the other is the discrete control part which gives the command when it receives messages from sensors. To steam boiler, it has got limits of water level such as minimal water level and maximal water level, and the water level influences the steam rate directly since little water cannot produce steam and too much water is likely to break the limit of steam rate and causes dangers. Furthermore, sensor collects the information of water level and steam rate, and then transfers the message to the controller. To the controller, it gives discrete commands to change the quantity of pumps so as to control the water volume and steam rate, and it has got some commands to reply the sensor fault and instrument fault. In addition, this paper only focuses on the basic steam boiler control system, but cannot tolerate the changes such as container volume expansion which is caused by heat etc. All the parameters are listed in Table 1.

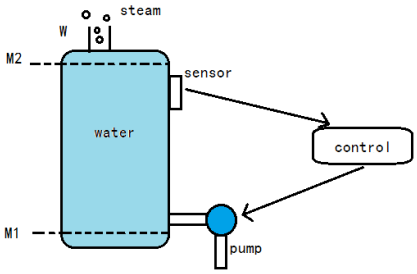


Fig. 1. Steam boiler control system

Table 1. Parameter list for Steam boiler control system

Parameter	Value
v_p	current speed of water
v_s	current speed of steam
V_w	current volume of water
e	volume of incoming water
z	volume change of water
M_1	minimal volume of water
M_2	maximal volume of water
C	capacity of steam boiler
W	limit of steam rate

To guarantee the safety of the control system, the specification should obey the next two rules. The current steam rate is always less than the highest rate **W** and the maximal volume of water cannot exceed the capacity of the steam boiler.

$$0 < v_s < W$$
$$0 < M_1 < M_2 < C$$

First of all, we consider the water level; **e** is the continuous variable which denotes the volume of incoming water; **minus_pump** and **add_pump** are events which control the quantity of pumps in order to adjust the steam rate. As the water volume is approaching the maximum, the controller is ready to make some pumps to stop, and the two events have time difference and sequence.

$$\mathrm{climb}(\mathrm{e}, M_2 - \mathrm{z}) \preccurlyeq \mathrm{clock}(\mathrm{minus\_pump})$$
$$\rho(\mathrm{climb}(\mathrm{e}, M_2 - \mathrm{z}), \mathrm{clock}(\mathrm{minus\_pump})) \leq \mathrm{z}/v_p$$
$$\mathrm{drop}(\mathrm{e}, M_1 + \mathrm{z}) \preccurlyeq \mathrm{clock}(\mathrm{add\_pump})$$
$$\rho(\mathrm{drop}(\mathrm{e}, M_1 + \mathrm{z}), \mathrm{clock}(\mathrm{add\_pump})) \leq \mathrm{z}/v_p$$
$$\mathrm{z} \leq \min\,(M_2 - V_w, V_w - M_1)/2$$

Since the water level can influence the steam rate, the following relationship is similar.

$$\mathrm{clock}(\mathrm{minus\_pump}) \preccurlyeq \mathrm{clock}(\mathrm{low\_steam})$$
$$\rho(\mathrm{clock}(minus\_\mathrm{pump}), \mathrm{clock}(\mathrm{low\_steam})) \leq \mathrm{z}/v_p$$
$$\mathrm{clock}(\mathrm{add\_pump}) \preccurlyeq \mathrm{clock}(\mathrm{high\_steam})$$
$$\rho(\mathrm{clock}(\mathrm{add\_pump}), \mathrm{clock}(\mathrm{high\_steam})) \leq \mathrm{z}/v_p$$

And it is vital that some couples of events in the equations above have noninterference.

$$\mathrm{clock}(\mathrm{add\_pump})[1] > 0$$
$$\mathrm{clock}(\mathrm{add\_pump}) \preccurlyeq \mathrm{clock}(\mathrm{minus\_pump}) \preccurlyeq \mathrm{clock}(\mathrm{add\_pump})'$$
$$\mathrm{clock}(add\_pump) \wedge \mathrm{clock}(\mathrm{minus\_pump}) = \emptyset$$
$$\mathrm{clock}(\mathrm{high\_steam}) \wedge \mathrm{clock}(\mathrm{low\_steam}) = \emptyset$$

There are also some timing issues in the system, since sensor and controller all need time to transfer the message and command. **sensor_delay** and **control_delay** denote the time consumed to transfer information.

$$\Big(\mathrm{clock}(high_{water}) \preccurlyeq \mathrm{clock}\big(\mathrm{sensor}_{\mathrm{delay}}\big)\Big) \preccurlyeq \mathrm{clock}\big(\mathrm{minus}_{\mathrm{pump}}\big)$$
$$\preccurlyeq \mathrm{clock}(\mathrm{control\_delay}) \preccurlyeq \mathrm{clock}(\mathrm{low\_steam})$$
$$\Big(\mathrm{clock}(low_{water}) \preccurlyeq \mathrm{clock}\big(\mathrm{sensor}_{\mathrm{delay}}\big)\Big) \preccurlyeq \mathrm{clock}\big(\mathrm{add}_{\mathrm{pump}}\big)$$
$$\preccurlyeq \mathrm{clock}(\mathrm{control\_delay}) \preccurlyeq \mathrm{clock}(\mathrm{high\_steam})$$

And no matter which part meets problem, the controller is fault-tolerant.

$$(\mathrm{clock}(\mathrm{sensor\_error}) \vee \mathrm{clock}(\mathrm{pump\_error})) \preccurlyeq \mathrm{clock}(\mathrm{control\_stop})$$
$$\preccurlyeq \mathrm{clock}(\mathrm{SYSTEM\_STOP})$$

In the equations, **e** denotes continuous speed change, and v_p denotes discrete speed at each clock unit. The continuous variable and discrete variable can be linked as below:

$$\dot{e} = v_p \; init \; v_{po}$$
$$v_p = e \; every \; c \; init \; v_{po}$$

4 Related Work

The UML profile for Modeling and Analysis of Real-Time and Embedded systems (Marte) has been adopted by the OMG earlier this year [7]. And Marte supersedes the UML Profile for Schedulability, Performance and Time (SPT) and extends the mainly untimed UML with several new constructs [8]. The Clock Constraint Specification Language (CCSL) [9] defines a set of time patterns between clocks that apply to infinitely many instant relations. A CCSL specification consists of clock declarations and conjunctions of clock relations between clock expressions. A clock expression defines a set of new clocks from existing ones. Most expressions deterministically define one single clock. In the paper of Lamport, concept of "happening before" defines an invariant partial order of the events in a distributed multiprocess system [10]. Since clocks range over the nonnegative reals, every nontrivial timed automaton has infinitely many states. If the clocks of a finitary real-time system are permitted to drift with constant, rational drift bounds, one obtains a finitary drifting-clock system. The representation of a closed finitary drifting-clock system as a graph annotated with constraints on drifting clocks is called an initialized rectangular automaton [11].

5 Conclusion

This paper applies a formal method called clock theory to the steam boiler control system, and describes the link between control events and continuous physical changes. For the specification of Cyber Physical Systems, it is tough to capture the dynamic change between discrete events and continuous time flows, so is the synchronization of all the subsystems in the whole system. Using clock to specify CPS can give more detailed description of every subsystem and more considerate observation of the time line and sequence of every event.

Acknowledgments. This work is supported by Shanghai Knowledge Service Platform Project (No.ZF1213), national high technology research and development program of China (No.2011AA010101), national basic research program of China (No.2011CB302904), the national science foundation of China under grant (No.61173046, No.61021004, No.61061130541, No.91118008), doctoral program foundation of institutions of higher education of China (No.20120076130003), national science foundation of Guangdong province under grant (No.S20110 10004905).

References

1. Kim, K.H.: Desirable Advances in Cyber-Physical System Software Engineering. In: 2010 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, pp. 2–4 (2010)
2. Kim, K.H.: Challenges and Future Directions of Cyber-Physical System Software. In: Proceedings of the 2010 34th Annual IEEE Computer Software and Applications Conference, pp. 10–13 (2010)

3. Derler, P., Lee, E.A., Vincentelli, A.S.: Modeling Cyber-Physical Systems. Proceedings of the IEEE 100(1), 13–28 (2012)
4. Chen, X., Liu, J., Mallet, F., Jin, Z.: Modeling Timing Requirements in Problem Frames Using CCSL. In: 2011 18th Asia-Pacific Software Engineering Conference, pp. 381–388 (2011)
5. He, J.: Link Continuous World with Discrete World. In: The 10th International Colloquium on Theoretical Aspects of Computing, Shanghai, China (unpublished, September 2013) (Keytalk)
6. Leeb, G., Lynch, N.: Proving safety properties of the steam boiler controller. In: Abrial, J.-R., Börger, E., Langmaack, H. (eds.) Formal Methods for Industrial Applications. LNCS, vol. 1165, pp. 318–338. Springer, Heidelberg (1996)
7. Object Management Group, UML Profile for MARTE, v1.0.formal (2009)
8. Object Management Group. UML Profile for schedulability, performance, and time specification. OMG document: formal(v1.1) (2005)
9. Mallet, F., DeAntoni, J., André, C., de Simone, R.: The clock constraint specification language for building timed causality models. Innovations in Systems and Software Engineering 6, 99–106 (2010)
10. Lamport, L.: Time, clocks, and the ordering of events in a distributed system. Commun ACM 21(7), 558–565 (1978)
11. Henzinger, T.A., Kopke, P.W., Puri, A., Varaiya, P.: What's decidable about hybrid automata? In: Proceedings of the 27th Annual Symposium on Theory of Computing, pp. 373–382. ACM Press (1995)

An Intelligent Dynamic Context-Aware System Using Fuzzy Semantic Language

Daehyun Kang[1], Jongsoo Sohn[2], Kyunglag Kwon[1], Bok-Gyu Joo[3], and In-Jeong Chung[1]

[1] Department of Computer and Information Science, Korea University
{internetkbs,helpnara,chung}@korea.ac.kr
[2] Service Strategy Team, Visual Display, Samsung Electronics
jongsoo.sohn@samsung.com
[3] Department of Computer and Information Communications, Hong-Ik University
bkjoo@hongik.ac.kr

Abstract. The prevalence of smart devices and the wireless Internet environment have enabled users to exploit environmental sensor data in a variety of fields. This has engendered various research issues in the development of context-awareness technology. In this paper, we propose a novel method where semantic web technology and the fuzzy concept are used to perform tasks that express and infer the user's dynamic context, in distributed heterogeneous computing environments. The proposed method expresses environmental information using numerical values, and converts them into fuzzy OWL. Then, we make inferences based on the user context, using FiRE, a fuzzy inference engine. The suggested method allows us to describe user context information in heterogeneous environments. Because we use fuzzy concepts to represent contextual information, we can easily express its degree or status.

Keywords: Context-aware computing, Fuzzy, Knowledge Representation, Inference, Fuzzy Web Ontology Language (OWL).

1 Introduction

For enhanced interaction with users in complex and distributed systems, developing dynamic context awareness systems becomes necessary that can recognize users, as well as information of the surrounding circumstances. Responding dynamically to changes in the application requirements, or the system itself, is also required [1]. With the advent of smart electronic devices, the problem of recognizing and expressing user context information, regardless of computer and language types, has emerged as an important task under the heterogeneous distributed processing system [2].

Since representing the environment that the user is in contact with the real world in crisp sets has some limitations, we introduce the fuzzy set as a more suitable means of representing the degree or status of the environment, than the crisp set [3]. For this purpose, we have chosen to use fuzzy Web Ontology Language (OWL) [4], a fusion

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_23,

of fuzzy concepts and the standard OWL to represent a user's dynamic context. In addition, it is used for the efficient description of a user's context, since it has the ability to represent the real context in a similar form to human thinking, independent of language and computer types, and infer new knowledge from the context data.

This paper suggests the following method. First, we represent user contacted environmental information with a numerical value and states, and describe it with OWL. Secondly, we transform the converted OWL context into fuzzy OWL [4]. Finally, we prove that automatic decision making of ambient environment is possible when using the fuzzy inference engine FiRE [5-6].

With the suggested method, we can describe the user context information in the ubiquitous computing environment. This method is effective in expressing both dynamic context information, and environmental status. We can also infer the user-contacted status of the environment. It is possible to enable this system to function automatically in compliance with the inferred state.

2 Related Works

A fuzzy OWL is one of the extended markup languages to represent a fuzzy set to OWL, which OWL itself does not provide [4, 7]. The fuzzy OWL provides a method to convert OWL into fuzzy OWL, and to describe membership functions that OWL is not able to. A fuzzy OWL uses the namespace 'fdl' to differentiate it from OWL. An element represented as a crisp set is described using OWL. Table 1 shows four principles to convert OWL into fuzzy OWL.

Table 1. Four principles to convert OWL into fuzzy OWL [4]

No	Principle
1	Every class in OWL is mapped into a corresponding fuzzy class in fuzzy OWL.
2	Every class subsumption or equivalence in OWL has a fuzzy subsumption or equivalence form in fuzzy OWL.
3	Every instance of class in OWL is mapped into a fuzzy constraint with restriction value, 1.
4	Every property in OWL has a primitive fuzzy property form in fuzzy OWL. Every instance of each property can be mapped into a fuzzy constraint with restriction value, 1.

The fuzzy OWL defines a namespace, Fuzzy Description Logic (FDL), and fuzzy constraints as shown in Table 2.

Table 2. Fuzzy constraints [4]

Rule	Fuzzy Constraints
$A(a) \geq n$	<fdl:individual fdl:name="a"> <fdl:membershipOf rdf:resource="#A" /> <fdl:moreOrEquivalent fdl:value=n /> </fdl:individual>
$A(a) \leq n$	<fdl:individual fdl:name="a"> <fdl:membershipOf rdf:resource="#A" /> <fdl:lessOrEquivalent fdl:value=n /> </fdl:individual>
…	…

3 Intelligent Context-Aware System Using Fuzzy OWL

Fig. 1 illustrates the overall architecture of the proposed system. The system is divided into three parts: (a) data collection, (b) context representation, and (c) context inference.

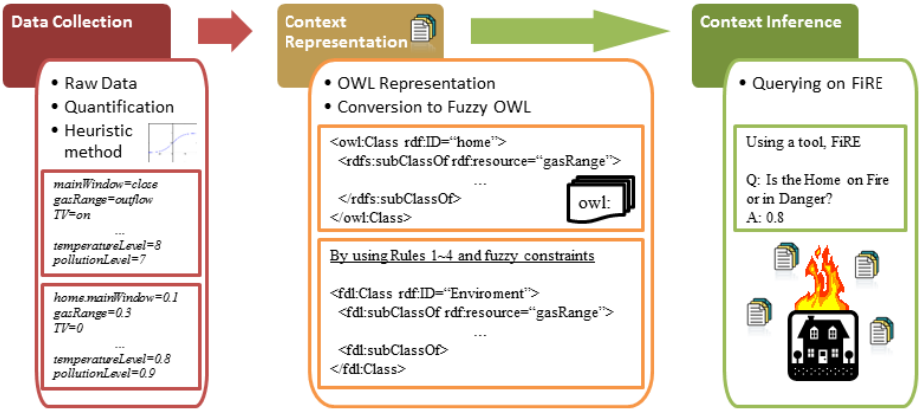


Fig. 1. Overall architecture of the suggested system

3.1 Data Collection

In the suggested system, data are automatically collected from a variety of sensors, such as door sensor, gas range sensor, and air pollution sensor, etc. in the form of raw or binary data. We then quantify the collected data using a heuristic method based on a fuzzy set. This is because decision-making in the computer is based on discrete data representation, such as set theory, and it is difficult for a computer to decide a situation from those context data. For example, we divide the level of temperature or pollution in a home into ten steps by using a fuzzy set. To be specific, the temperature below 0℃ is level one, temperature between 0℃ and 10℃ is level two, and so on. In

the same manner, the value of indoor pollution between 0ppm and 50ppm is level one, the value of indoor pollution between 51ppm and 100ppm is level two, and so on.

Next, we normalize the quantified data as a real number between 0 and 1. This is used for a weight value in the fuzzy inference formula. For instance, the fact that the value of air pollution is very high is more related to a fiery, dangerous situation than the fact that the door is open, or the TV is on. Thus, the former context has a higher value than the latter one.

3.2 Context Representation

Based on the normalized data and their relations to a situation, we describe the situation as fuzzy OWL. There are two steps to describe a context as fuzzy OWL.

First, we convert factors into a context, which can be represented by a general set in OWL. A factor can be one of every concrete surrounding element to describe a situation, such as doors, TV, gas range, windows, etc. Second, we transform the OWL contexts into fuzzy OWL using constraints, as shown in Table 2. We specify an '*fdl*' namespace to describe the context as fuzzy OWL, otherwise an '*owl*' one. If a factor has an weight value x between 0 and 1, we represent the value as a property '*fdl:value* = x' in fuzzy OWL, as follows.

<*fdl*:Class rdf:about = "temperatureLevel" />
  <*fdl*:moreOrEquivalent *fdl:value* = "*0.8*" />

3.3 Context Inference

To infer a situation based on the described fuzzy OWL, we classify each factor or element by its characteristics, and define a set of inference rules. If a situation happens at home, for example, we can classify the contexts into several classes, such as electronic home appliances, environmental information, family, location, and so on. Each class has atoms; for example, a class *Family* has atoms, such as father, mother, brother, sister, daughter, etc., and a class *Environment* has atoms, such as temperature, air pollution, and so on.

Afterwards, we define inference rules using a conjunction operator (∧). We assume a home that is on fire as an example. The situation includes the facts that the level of temperature and air pollution is high, and whether the gas range is turned on or not. We define an inference rule for this fiery situation, as shown in Equation (1).

$$\text{TemperatureLevel}(?k) \times 0.7 \wedge \text{AirPollutionLevel}(?k) \times 0.9 \wedge \text{GasRange}(?k) \rightarrow \text{OnFire}(?k) \quad (1)$$

In Equation (1), each value represents the weight value of each factor that occurs in the situation of a house on fire. Through the definition of inference rule, we can infer the value of context OnFire(?k) using each parameter k. The parameter k includes a set of properties in each area such as kitchen, living room, and bathroom.

4 Implementation Examples

4.1 Context Representation

In this section, we demonstrate a whole process where the environmental data received from home network sensors are represented in context. We assume that the following environmental data are collected, as shown in the left part of Table 3.

Table 3. Collected environmental data, and their conversion to context data

Collected environmental data	Converted context data
home.mainWindow = closed	home.mainWindow = 0.1
home.gasRange.gas = outflow	home.gasRange.gas = 0.3
home.gasRange.fire = on	home.gasRange.fire = 0.3
home.TV = on	home.TV = 0
home.family.daughter = in	home.family.daughter = 0
home.temperature = 72.3	home.temperatureLevel = 0.8
home.air.pollutionLevel = 8	home.air.pollutionLevel = 0.9

The data describe that gas is flowing out from a gas range, the indoor temperature is high enough, and the level of indoor pollution is high. We can regard this situation as the scene of a fire. Therefore, we convert these environmental context data into real number values between 0 and 1 with a heuristic method before expressing them using the OWL, as described in the right part of Table 3.

We then represent context information in the form of fuzzy OWL, based on the conversion rules proposed in [4]. Fig. 2 shows the fuzzy OWL context representation of the collected environmental context data. If we use fuzzy sets with weighted values between 0 and 1, we can describe the weighted value of *fdl:value* = "0.9", as shown in the middle of Fig. 2.

```
<?xml version="1.0" encoding="UTF-8"?>
- <fdl:individual xmlns:rdf="http://www.w3.org/2000/01/rdf-
schema#" xmlns:fdl="http://iis.korea.ac.kr/2012/fdl"
fdl:name="fire">
  - <fdl:membershipOf>
      - <fdl:Restriction>
          <fdl:onProperty rdf:resource="#homeContext"/>
        - <fdl:someValuesFrom>
            - <fdl:Class>
                - <fdl:unionOf rdf:parseType="Collection">
                    <fdl:Class
                      rdf:about="#homeTemperature"/>
                    <fdl:moreOrEquivalent fdl:value="0.8"/>
                    <fdl:Class rdf:about="#airPolluteLevel"/>
                    <fdl:moreOrEquivalent fdl:value="0.9"/>
                  </fdl:unionOf>
              </fdl:Class>
          </fdl:someValuesFrom>
      </fdl:Restriction>
  </fdl:membershipOf>
</fdl:individual>
```

Fig. 2. Fuzzy OWL-based context representation

4.2 Contextual Inference

We use an inference engine called FiRE to infer the user context and its corresponding services. The FiRE is based on Fuzzy Description Logic, and f-SHIN [8] provides sufficient grammars to use Description Logic [9], as shown in Table 4.

Table 4. Examples of Inference rules

Inference Rule #1: OnFire(?k)	temperatureLevel(?*k*) × 0.7 ∧ AirPollutionLevel(?*k*) × 0.9 ∧ GasRange(?*k*) → OnFire(?*k*)
Inference Rule #2: Danger(?k, ?f)	OnFire(?*k*) × 0.5 ⋀ Home(?*k*, ?*f*) → Danger(?*k*, ?*f*)

In order to utilize the FiRE, we describe declarations of atomic-concept rules, an axiom, and an ABox. In Table 5, a part of 'atomic-concepts' represents an enumeration of each element in a set, and 'roles' describe the relationships between individuals. Table 6 specifies the axioms to demonstrate how it infers a situation where a fire broke out in the FiRE. Table 7 shows a defined ABox, which describes the converted values using fuzzy OWL. The values can be used by the FiRE.

Fig. 3 shows the inference results for 'Home Danger' obtained from Table 7. The output value 0.8 means the degree of danger is 0.8 for the given situation in Table 7.

Table 5. Atomic-concepts rules and individuals

(signature :atomic-concepts (person mother father daughter Temperature AirPollution gasRange) :roles ((has-gender :*transitive* t)	(has-descendant :*inverse* t) (has-child :*inverse* has-descendant) (has-sibling) (has-degree)) :individuals (home))

Table 6. Axioms

(implies person (and human (some has-gender (or female male))))
(equivalent daughter (and woman (some has-sibling person)))
(equivalent OnFire (and Temperature airPollution GasRange))
(equivalent Danger (and OnFire (some inHome person)))

Table 7. An example of an ABox

(instance Home temperatureLevel >= 0.8)	(instance Home GasRange >= 1.0)
(instance Home airPollutionLevel >= 0.9)	(related Home daughter inHome)

Fig. 3. Query result for 'Home Danger'

5 Conclusions and Future Works

In this paper, we expressed user context information using a fuzzy extended language version of OWL, i.e. fuzzy OWL. Fuzzy OWL is suitable for expressing the user context necessary in a ubiquitous computing environment, while it also provides a basis for the effective representation of crisp sets and fuzzy sets.

The method proposed in this paper uses an ontology language in a ubiquitous computing environment to describe user's dynamic context information, independent of computer types and languages. Using the fuzzy concept, we can express problems and contexts in the real world, which are difficult to represent using the binary values of 0 and 1. We also provide a foundation for making further inferences in real world situations. When constructing an intelligent context awareness system with user context information, the result may vary depending on how we apply and implement the inference rules in the knowledge base.

In future, we will provide more examples of real world applications, and implement an inference system using the fuzzy ontology that we have created. Using our complete inference system, we can construct an intelligent dynamic context awareness system for different types of languages and computers.

Acknowledgment. This research was partially supported by Korea University.

References

1. Dey, A.K.: Providing Architectural Support for Building Context Aware Applications. Georgia Institute of Technology (2000)
2. Stoilos, G., Stamou, G., Pan, J.Z.: Fuzzy Reasoning Extensions. In: Knowledge Web Consortium (2007)
3. Chen, H., Wu, Z.: Semantic Web Meets Computational Intelligence: State of the Art and Perspectives. IEEE Computational Intelligence Magazine 7, 67–74 (2012)
4. Gao, M., Liu, C.: Extending OWL by Fuzzy Description Logic. In: 17th IEEE International Conference on Tools with Artificial Intelligence (2005)
5. Simou, N., Kollias, S.: FiRE: A Fuzzy Reasoning Engine for Impecise Knowledge. In: K-Space PhD Students Workshop (2007)
6. Simou, N., Stoilos, G., Stamou, G.: Storing and Querying Fuzzy Knowledge in the Semantic Web Using FiRE. In: Bobillo, F., Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2008-2010/UniDL 2010. LNCS, vol. 7123, pp. 158–176. Springer, Heidelberg (2013)
7. Huang, C., Lo, C., Chao, K.: Reaching consensus: A moderated fuzzy web services discovery method. Information and Software Technology 48, 410–423 (2006)
8. Stoilos, G., Stamou, G., Tzouvaras, V.: The fuzzy description logic f-SHIN. In: Proc. of the International Workshop on Uncertainty Reasoning for the Semantic Web (2005)
9. Pan, J.Z., Stamou, G., Stoilos, G., Thomas, E.: Fuzzy querying over fuzzy-DL-Lite. In: 17th International World-Wide-Web Conference, Beijing (2008)

Efficient Data Monitoring in Sensor Networks Using Spatial Correlation

Jun-Ki Min

School of Computer Science and Engineering,
Korea University of Technology and Education,
Byeoncheon-myeon, Cheonan, ChungNam, Republic of Korea, 330-708
jkmin@koreatech.ac.kr

Abstract. In order to reduce r the energy consumption of sensors, we present an approximate data gathering technique, called CMOS, based on the Kalman filter. The goal of CMOS is to efficiently obtain the sensor readings within a certain error bound. In our approach, spatially close sensors are grouped as a cluster. Since a cluster header generates approximate readings of member nodes, a user query can be answered efficiently using the cluster headers. Our simulation results with synthetic data demonstrate the efficiency and accuracy of our proposed technique.

Keywords: sensor network, data monitoring, Kalman filter.

1 Introduction

Energy preservation is a major research issue in the sensor network field since it directly impacts the lifetime of a network [5, 4, 7]. Thus, to reduce the communication overhead, in-network approximation techniques [2, 4, 6] have been proposed.

Generally, in approximate techniques, each sensor estimates a sensor reading v as ^v. If the difference of ^v and the actual reading v is greater than a user specific threshold ϵ (i.e., $|\hat{v}-v| > \epsilon$), each sensor transmits v to the base station. In the base station, a mirror model is maintained to predict a sensor reading of each sensor. Thus, if a sensor node does not send a sensor reading, the base station obtains an approximated sensor reading using the mirror model. For most techniques of this approach, each sensor estimates its reading independently. A sensor's neighbor refers to any other sensor that is within its communication distance. In the sensor field, the spatial correlations such that the change patterns of two neighbors' sensor readings are the same or similar occur. Therefore, in this paper, we propose CMOS, a Cluster based MOnitoring technique for Sensor networks utilizing the spatial correlation. To estimate sensor readings, CMOS utilizes the Kalman filter which requires the previously predicted future value and the current measure value to predict a future value.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_24,

2 Related Work

Most applications of sensor networks do not require highly accurate data. Therefore, some approximated data gathering techniques were introduced. To estimate sensor readings, Tulone and Madden devised PAQ [8] based on an autoregressive model (AR). Particularly, in [8], a dynamic AR(3) model is used in which a future reading is predicted using recent three readings with the following equation, $X(t) = \alpha X(t-1) + \beta X(t-2) + \gamma X(t-3) + b(\omega)N(0,1)$, where $b(\omega)N(0,1)$ represents the Gaussian white noise of zero mean and standard deviation $b(\omega)$. But, PAQ requires a long learning phase to build the proper coefficients α, β, and γ of AR(3) model.

Jain et al. suggested Dual Kalman Filter [2] which is based on the Kalman filter. In addition, recently, Min and Chung proposed EDGES [6] based on a variant of theKalman filter, i.e., multi-model Kalman filter. In these approaches, each sensor estimates its readings independently with its own model.

Unlike the other techniques (PAQ, the Dual Kalman filter, and EDGES), we consider the spatial correlations such that the change patterns of sensor readings of the neighbor sensors are the same or similar. To utilize this correlation, in our work, sensors in WSN are grouped as a cluster and each sensor has several Kalman filters each of which serves a different purpose (see details in Section 4).

In [4], the snapshot query approach was introduced. In this work, nodes can coordinate with their neighbors and elect a small set of representative nodes among themselves. In order to maintain the representative nodes, the authors assume that each node knows the values of its neighbors. For this, sensors periodically broadcast their readings to their neighbors as heartbeat messages. It wastes lots of energy since each node should receive the data of its neighbors. Also, since a representative node does not know its non-representative nodes' data values within an interval of the non-representatives' periodic data sending, the error bound cannot be guaranteed.

3 Preliminary

The Kalman filter [3] consists of mathematical equations that estimate the internal states of a system using a predicate-corrector type estimator as shown in Figure 1.

In the Kalman filter, the system model is represented by the following equations.

$$x_k = Fx_{k-1} + w_{k-1} \tag{1}$$

$$z_k = Hx_k + v_k \tag{2}$$

Equation (1) represents a process model that shows the transformation of the process state. Let $x_k \in \Re^n$ be the state of a process. F is the $n \times n$ state transition matrix relating the state x_k and x_{k-1}. Equation (2) represents a measurement model that describes the relationship between the process state and the measurement $z_k \in \Re^m$. H is the $m \times n$ matrix relating the state to the measurement. $w_k \in \Re^n$ and $v_k \in \Re^m$ represent the process noise and measurement noise, respectively.

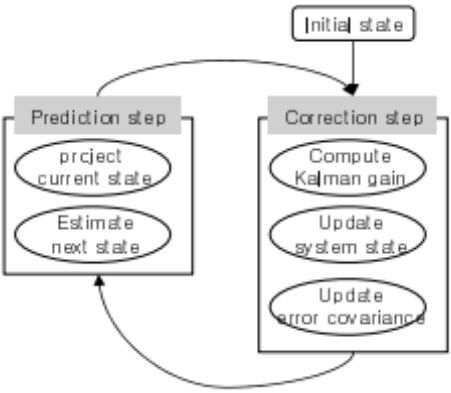


Fig. 1. Recursive cycle of the Kalman filter

To estimate the process state x, the Kalman filter uses the estimators $\hat{x}_k$ and $\hat{x}\text{-}_k$. $\hat{x}_k$ is called a posteriori state estimate at time k given measurement z_k. And $\hat{x}\text{-}_k$ is called a priori state estimate at time k using a previously estimated posterior state $\hat{x}_{k-1}$. $\hat{x}\text{-}_k$ and $\hat{x}_k$ are computed by the following equations:

$$\hat{x}\text{-}_k = F\,\hat{x}_{k-1} \tag{3}$$

$$\hat{x}_k = \hat{x}\text{-}_k + K_k(z_k - H\,\hat{x}\text{-}_k) \tag{4}$$

In the discrete Kalman filter, by using the equation (3), the prediction of a future value is conducted. And, by using the equation (4), the correction of an estimated value (i.e., measurement update) is performed. In equation (4), the $n \times m$ matrix K_k is called Kalman gain. As presented in the above equations, the Kalman filter does not store the previous data set nor reprocess stored data if a new measurement becomes available. In other words, to predict a future value at time k, the Kalman filter only requires the previously predicted future value at time $k-1$ and a measurement value at time k [6].

4 CMOS

In CMOS, sensor nodes in a network are grouped into clusters and each cluster elects a cluster header. A cluster header communicates with the base station through multi-hop routing. The maximum distance between a cluster header and its member nodes is c (i.e., a hop distance). Since member nodes and the respective cluster header are located closely, the spatial correlation such that the changing patterns of sensor readings of the neighbor sensors are the same or similar occurs.

As mentioned earlier, CMOS estimates sensor readings using the Kalman filter. For the data model of the Kalman filter, we use the uniform velocity model. In CMOS, $x_k = [v_k, r_k]^T$ is used as a process state where v_k is a value and r_k is the rate of change (i.e., velocity) of v_k. Under the uniform velocity model, $v_k = v_{k-1} + r_{k-1}\Delta t$ and r_k

$= r_{k-1}$ where Δt is an elapse time between k and $k - 1$. Thus, we make a state transition matrix F as follows:

$$F = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}$$

Then, let the measurement of a value (i.e., the actual value) be $z_k \in \Re$. The state measurement matrix H is represented as follows: $H = [1\ 0]$

In CMOS, a cluster header CH estimates its reading v_{ch} based on the process model and measurement model (i.e., F and H). If the difference of the actual value v_{ch} and the estimate value $\hat{v}_{ch}$ is greater than ϵ (i.e., $|v_{ch} - \hat{v}_{ch}| > \epsilon$), CH will report v_{ch} to the base station. Otherwise, the base station can obtain $\hat{v}_{ch}$ as a report value using the Kalman filter KF_{CH} for CH. A member node mi maintains the difference dmi between its reading v_{mi} and the cluster header's report value v_{chrep} (i.e., $d_{mi} = v_{mi} - v_{chrep}$) using the Kalman filter KF_{dmi} under the uniform velocity model. As mentioned above, the cluster header CH estimates v_{ch} as $\hat{v}_{ch}$. Thus, in a member node, the cluster header's report value v_{chrep} is $\hat{v}_{ch}$ if the cluster header does not broadcast v_{ch} (i.e.,$|v_{ch} - \hat{v}_{ch}| <= \epsilon$). Otherwise,v_{chrep} is v_{ch}.

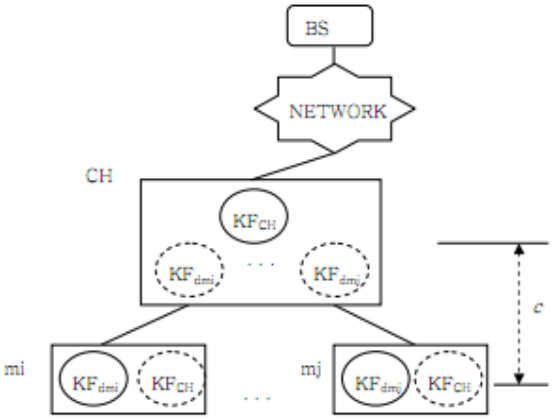


Fig. 2. An architecture of a cluster

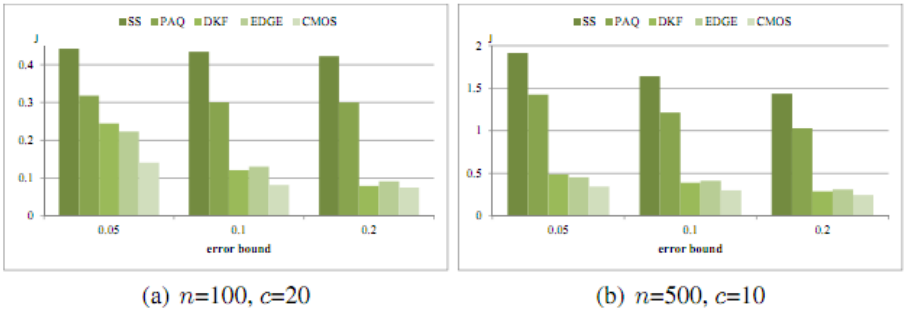
The basic architecture of a cluster in CMOS is presented in Figure 2. As shown in Figure 2, CH has the Kalman filter KF_{CH} in order to estimate its reading vch. Each member node has the mirror KF_{CH} represented as a dotted circle in Figure 2.

Each member node mi also has the Kalman filter KF_{dmi} in order to estimate the difference d_{mi} of its own reading and CH's reading. CH has the mirror KF_{dmi} s for its member nodes. In addition, the base station keeps the information of the clusters including the Kalman filters for cluster headers and their members. Thus, the base station can estimate properly sensor readings which are measured in a cluster properly.

5 Performance Study

In this section, we demonstrate the efficiency of our proposed method, CMOS. We perform simulations to compare the performance of CMOS with snapshot approach (SS) [4], PAQ [8], Dual Kalman filter (DKF) [2] and EDGES [6] on the synthetic data. The sensor network consists of 100 and 500 sensors, randomly located in the [0,100) × [0,100) two dimensional-sensing field. In addition, we locate the base station at the center of the sensing field for all data sets. The communication distance c on the synthetic data is 20 or 10. To measure the energy consumption in diverse environments, we use three error bounds (ϵ), 0.2, 0.1, and 0.05.

For the synthetic data, we make the Wave data set. For the Wave data set, we assign a value in the range [0.0... 50.0] to a location in the [0, 100) space using the SIN function. We set the values to the two-dimensional space using the assigned values, where locations with the same x-coordinates have the same value. Then, we simulate the wave passing as the vertical shift from left to right.



(a) n=100, c=20 (b) n=500, c=10

Fig. 3. Average energy consumption on Wave data

Figure 3 shows the energy consumption on the Wave data, averaged over sensors. Generally, as the error bound increases, the energy consumption decreases since the number of data transmissions decreases. As shown in Figure 3, DKF and EDGES show similar performances. CMOS shows the best performance over all cases. This result indicates that our approximate data monitoring technique based on spatial correlation is effective over all cases. In addition, as shown in Figure 3-(a) and (b), when the number of sensor nodes increases, the performance gap between PAQ and CMOS increases. This results show that CMOS is more scalable than PAQ.

6 Conclusion

In this paper, we propose an efficient cluster based monitoring technique called CMOS. In CMOS, sensors in networks are grouped into clusters. The cluster header in a cluster predicts its reading and member nodes predict the differences of their readings and the cluster header's reading using the Kalman filters. Since each node

keeps the mirror Kalman filter for the counterpart, a cluster header (member) node can estimate the reading of a member (header) without data transmission.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1B3003060).

References

1. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocolarchitecture for wireless microsensor networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)
2. Jain, A., Chang, E.Y., Wang, Y.-F.: Adaptive stream resource management using kalmanfilters. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 11–22 (June 2004)
3. Kalman, R.E.: A new approach to linear filtering and prediction problem. Transactions of ASME Journal of Basic Engineering 82, 34–45 (1960)
4. Kotidis, Y.: Snapshot queries: Towards data-centric sensor networks. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE), pp. 131–142 (April 2005)
5. Madden, S., Franklin, M.J., Hellerstein, J.M., Hong, W.: Tag: A tiny aggregation service forad-hoc sensor networks. In: 5th Symposium on Operating System Design and Implementation (OSDI) (December 2002)
6. Min, J.-K., Chung, C.-W.: Edges: Efficient data gathering in sensor networks using temporaland spatial correlations. Journal of Systems and Software 25(5), 933–944 (2010)
7. Stern, M., Bohm, K., Buchmann, E.: Processing continuous join queries in sensor networks: a filtering approach. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, pp. 267–278 (2010)
8. Tulone, D., Madden, S.: PAQ: Time series forecasting for approximate query answering in sensor networks. In: Römer, K., Karl, H., Mattern, F. (eds.) EWSN 2006. LNCS, vol. 3868, pp. 21–37. Springer, Heidelberg (2006)

Power-Time Tradeoff of Parallel Execution on Multi-core Platforms

Sungju Lee, Heegon Kim, and Yongwha Chung

Department of Computer Information Science, Korea University, Sejong, Korea
{peacfeel,khg86,ychungy)@korea.ac.kr

Abstract. It is anticipated that high-performance handheld multi-core devices will be used as WSN processing nodes in the near future. Reducing execution time by deploying parallel applications on multi-core platforms comes at the cost of increasing power consumption compared to using fewer cores. This paper focuses on such tradeoff between power consumption and execution time and subsequently achieves maximal energy saving when executing applications in parallel. Based on the experiments on a multi-core platform, we can verify that parallel execution with frequency scaling is an effective approach at the application level in order to reduce energy consumption.

Keywords: multi-core platform, energy saving, parallel application.

1 Introduction

With the advances made in multi-core technology, multiprocessing techniques at the system software level have been used in order to reduce energy consumption [1]. However, it has been rarely reported that parallel processing techniques at the application level can reduce the application energy consumption. Since a multi-core microprocessor has been widely generalized in the mobile handheld devices [2], we need to verify whether or not parallel processing on multi-core based mobile platforms is an effective way for energy saving.

Traditionally, parallel processing has been widely used to reduce the application execution time [3,4]. However, since using more cores causes higher power consumption compared to using fewer cores, it is important to quantitatively evaluate the tradeoff between execution time and power consumption of parallelized applications.

The previous studies such as [5,6], conducted by the computer architecture community were targeted for designing general-purpose processors which can be applied to several applications. That is, processor vendors provide several levels of frequency settings and several numbers of cores, and it is a user's role to determine the optimal configuration for his/her application. Accordingly, it is a possible way to reduce the power consumption by controlling the frequency and the number of cores based on commercial multi-cores platforms.

In this paper, we first evaluate the machine characteristics (*i.e.*, the growth rate of power consumption with the increased number of cores and with the increased

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_25,

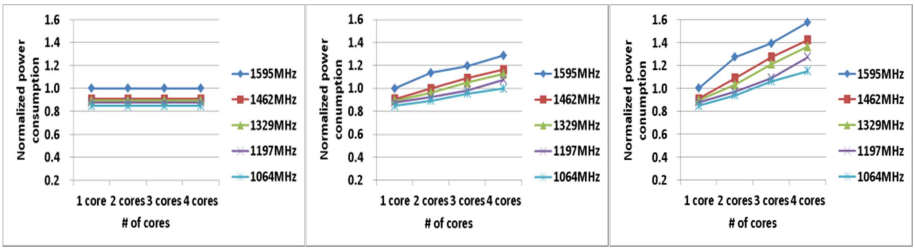
frequency as well) for given multi-core platform. Then, we derive the condition for energy saving with parallel processing by analyzing the application characteristics (*i.e.*, the fraction of sequential work and the parallelization overhead). Based on the experimental results, we confirmed that the combined technique of parallel processing and frequency scaling can significantly reduce (by up to 60%) the energy consumption of JPEG encoding while reducing the total encoding time, which clearly allows multimedia applications to satisfy their temporal constraints.

The rest of the paper is structured as follows: Section 2 explains the proposed approach for the analysis of machine characteristics and application characteristics, and the optimization of system configuration. Sections 3 and 4 describe the experimental results and conclusions, respectively.

2 Proposed Approach

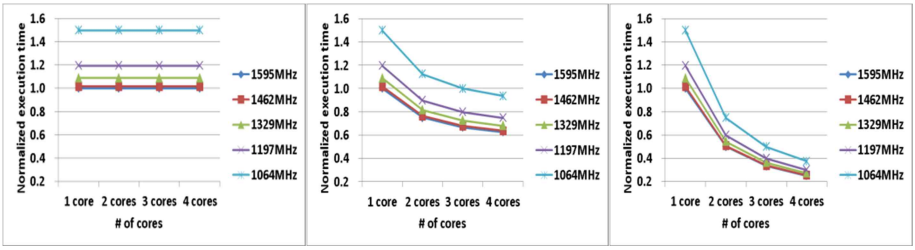
First, to understand the machine's and application's characteristics, we measured the power consumption, execution time, and the energy consumption of parallelized AES-CBC (*i.e.*, Advanced Encryption Standard with Channing Block Cipher and 0% parallelism), AES-CCM (*i.e.*, Advanced Encryption Standard with Counter with CBC-MAC and 50% parallelism) and AES-CTR (*i.e.*, Advanced Encryption Standard with Counter 100% parallelism) [7] with the *Pthread* library [8] as examples of test applications on the Intel i7 multi-core processors. The AES-CTR problem has no data dependency and is easily parallelized. In contrast, AES-CCM has 50% data dependency, and AES-CBC has 100% data dependency. According to Amdahl's law, the maximum speedup (with a 4-core processor) of AES-CTR and AES-CCM are 4 and 2, respectively.

Fig. 1 and Fig. 2 show the power consumption and execution time of the test applications with 0%, 50%, 100% parallelism on multi-core processors, with various frequencies and numbers of cores. The power consumption, the execution time were normalized based on the case with a single core and maximum frequency. As shown in Fig. 1, if an application's parallelism is more than 0, the power consumption increased with increased number of cores. In contrast, the power consumption increased with increased frequency regardless of application's parallelism.



Parallelism of 0% Parallelism of 50% Parallelism of 100%

Fig. 1. Normalized power consumption with different frequency and number of cores



Parallelism of 0% Parallelism of 50% Parallelism of 100%

Fig. 2. Normalized execution time with different frequency and number of cores

Although an application with less parallelism requires less power consumption, it may consume more energy due to greater execution time. Fig. 2 shows the execution time of AES-CBC, AES-CCM, and AES-CTR on 1, 2, 3, and 4 cores. AES-CBC (0% parallelism) can be performed with increased number of the cores, but both the power consumption and the execution time are always constant (see Fig. 1 and Fig. 2). In contrast, as we increase the number of cores in AES-CTR (100% parallelism), the execution time decreases while the power consumption increases. To improve the energy efficiency, we need a collective analysis of the machine and application characteristics.

To collectively analyze the machine and application characteristics, we analyze the relationship between the application/machine and the energy consumption. The power consumption and the execution time depend on the characteristics of the machine and the application. Thus, we can represent the energy consumption E by Equation (1) with power consumption W and execution time T:

$$E = W \times T \tag{1}$$

To analyze the power consumption and the execution time with an application's parallelism, we denote the application's parallelism as p_{app}, where $0 \leq p_{app} \leq 1$.The application's parallelism (*i.e.*, p_{app}), frequency (*i.e.*, f), and number of cores (*i.e.*, n) sensitively affect the energy consumption of a processor as shown in Fig. 3. Thus, the energy consumption is represented as Equation (2), where f is the frequency and n is the number of cores. To reduce the energy consumption, we need to set the optimal f and n with a prediction of the energy consumption from the given application and machine characteristics.

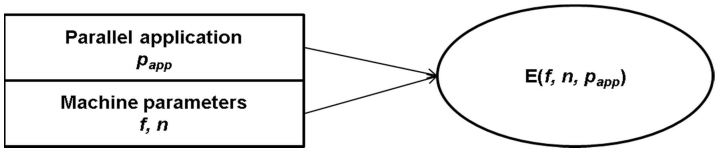


Fig. 3. The relationship between application/machine characteristics and the energy consumption

$$E(f, n, p_{app}) = W(f, n, p_{app}) \times T(f, n, p_{app}) \quad (2)$$

The commercial multi-core processor has different characteristics by the design of the hardware architecture. For example, in Intel's multi-core architecture, L2 cache is shared by two cores. In AMD's multi-core architecture, L2 cache is allocated per core. Additionally, according to the service requirement, various hardware components(*i.e.*, memory, hard disk, IO devices, etc.) can be configured. Since the characteristics of the power consumption and execution time of the commercial multi-core processor depend on the design of the hardware architecture, it is difficult to generalize the characteristics of the power consumption and execution time. Therefore, to analyze the machine's characteristics, the power consumption and execution time need to be measured at least once.

The power consumption can be measured with an application having 100% parallelism (*i.e.*, AES-CTR). With an increased number of cores, the power consumption is also increased. We can also find that the power consumption depends on the number of cores. Thus, when the combination of application and machine characteristics are given, we can analyze the application's parallelism. We can predict the power consumption by using Equation (3) with the measured results. We denote the power consumption of the sequential portion of the application with *1* core as $W_{sequential}(f, 1)$ and the power consumption of the parallel portion of the application with n cores $W_{parallel}(f, n)$. Note that, W is average watt during the execution time.

$$W(f, n, p_{app}) \approx W_{sequential}(f, 1)\times(1-p_{app}) + W_{parallel}(f, n)\times(p_{app}) \quad (3)$$

The parallelism is also application dependent and the speedup can vary greatly even for different versions of the same software. Therefore, we analyze not only the speedup for parallelism of an application, but also the speedup of increased frequency.

The total execution time (*i.e.*, $T(f, n, p_{app})$, with various numbers of cores can be predicted using Equation (4). $W_{sequential}(f, 1)$ and $T_{sequential}(f, 1)$ represent the power consumption and the execution time of the sequential portion of the application, respectively. We denote the execution time of the sequential portion of the application with *1* core as $T_{sequential}(f, 1)$ and the execution time of the parallel portion of the application with n cores $T_{parallel}(f, n)$. Note that, if we denote the execution time of the parallel portion of the application with *1* core as $T_{parallel}(f, 1)$, then $T_{parallel}(f, n)$ is not equal to $T_{parallel}(f, 1)/n$ in a strict sense, due to the pthread overhead. However, $T_{parallel}(f, n)$ can be approximately equal to $T_{parallel}(f, 1)/n$, with a careful parallelization:

$$T(f, n\, p_{app}) \approx T_{sequential}(f, 1)\times(1-p_{app}) + T_{parallel}(f, 1)/n \times(p_{app}) \quad (4)$$

In this paper, we propose a greedy approach to find the local optimal parameters for the energy efficiency in transmitting image/video data without compromising image/video quality. Algorithm 1 shows the procedure to find the local optimal frequency f and the number of cores n by using a greedy approach. Although Algorithm 1 provides the local optimal parameters instead of global optimal parameters, it can reduce the energy consumption similarly to global optimal parameters, and provide more fast speed than brute-force search.

Algorithm 1. Finding Local Optimal Machine Parameters.

given the environment parameter
  p_{app}← application's parallelism
set the default parameters
  f ← maximum frequency and n ← 1 core
do {
  calculate $E(f, n, p_{app})$
  if (n_next is not last level) {
    n_next ← next increased level
    calculate $E(f, n\_next, p_{app})$}
  if (f_next is not last level) {
    f_next← next decreased level
    calculate $E(f\_next, n, p_{app})$}
  if ($E(f, n\_next, p_{app})<E(f, n, p_{app})$) n←n_next
  if ($E(f\_next, n, p_{app})<E(f, n\_next, p_{app})$) f←f_next
} while (($E(f, n, p_{app})<E(f, n\_next, p_{app})$ AND $E(f, n, p_{app})<E(f\_next, n, p_{app})$)
f_opt←f and n_opt←n // found local optimal frequency and number of cores

3 Experimental Result

To verify our analysis, we estimated the speedup of JPEG encoding that can be used for battery-operated portable devices equipped with a multi-core processor. We parallelized JPEG encoding and derived the optimum parameters in order to reduce the energy consumption. Since the fraction of sequential work (*i.e.*, reading the input image) of JPEG encoding was 3%, the normalized (w.r.t. 1-thread time) execution times with 2, 3, 4 cores were estimated to 0.52, 0.35 and 0.27, respectively. To check whether these parameters can achieve the minimum energy, we parallelized JPEG encoding with data domain decomposition [4] using Pthread library [8]. That is, the whole image data was partitioned into subimages, and each subimage was assigned to each thread. Then, we measured the time and power consumption. Table 1 summarizes the experimental environments(*i.e.*, given application and machine characteristics).

Table 1. Platforms spec. with Intel i7 processors.

Application Characteristics	Application	JPEG release version 8c
	Application parallelism	0.96
Machine Characteristics	Processor	Intel i7 720QM
	Frequency range	1.0GHz~1.5GHz
	Frequency step	133MHz
	The maximum # of cores	4

Table 2. Estimated/Measured normalized execution time and energy consumption for JPEG

Estimated normalized energy consumption for JPEG.

	1064MHz	1197MHz	1329MHz	1462MHz	1596 MHz
1 core	127%	117%	108%	99%	100%
2 cores	74%	67%	63%	61%	65%
3 cores	55%	51%	49%	49%	51%
4 cores	46%	44%	43%	**42%**	44%

Measured normalized energy consumption for JPEG.

	1064MHz	1197MHz	1329MHz	1462MHz	1596 MHz
1core	127%	117%	107%	99%	100%
2 cores	72%	66%	61%	60%	63%
3 cores	52%	49%	46%	47%	49%
4 cores	42%	41%	40%	**39%**	41%

Then, the estimated/measured energy consumptions with different frequency and number of cores are summarized in Table 1. Note that, the aforementioned machine characteristics need to be measured only once and the application characteristics can be estimated straight forwardly by using the Amdahl's law. As shown in Table 1, we confirmed that our estimation was accurate enough to locate the appropriate parameters for energy-efficient parallel processing.

4 Conclusion

In this paper, we proposed an analysis framework to save energy by parallelizing applications and executing them on multi-core platforms with frequency scaling. We first analyzed the machine characteristics, and then combined them to the application characteristics in order to derive the optimum number of threads and frequency level. That is, our analysis provides a practical guideline in finding an energy-efficient solution for portable devices at the application level. Experimental results show that the energy consumption of an application can be reduced significantly (by up to 60%) compared to the sequential execution with the maximum frequency.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2043679) and supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-S1A5B5A01- 2012R1A6A3A01040440).

References

1. Huang, X., Li, K., Li, R.: A Energy Efficient Scheduling Base on Dynamic Voltage and Frequency Scaling for Multi-core Embedded Real-Time System. In: Hua, A., Chang, S.-L. (eds.) ICA3PP 2009. LNCS, vol. 5574, pp. 137–145. Springer, Heidelberg (2009)
2. Levy, M., Conte, T.: Embedded Multicore Processors and Systems. IEEE Micro. 29(3), 7–9 (2009)

3. Kumar, V., Grama, A., Gupta, A., Karypis, G.: Introduction to Parallel Computing – Design and Analysis of Algorithms. The Benjamin/Cummings Pub. Co. Inc. (1993)
4. Castells-Rufas, D., Joven, J., Carrabina, J.: Scalability of a Parallel JPEG Encoder on Shared Memory Architectures. In: ICPP 2010 Pub., pp. 13–16 (2010)
5. Li, J., Martınez, J.: Dynamic Power-Performance Adaptation of Parallel Computation on Chip Multiprocessors. In: International Symposium on High-Performance Computer Architecture, pp. 77–87 (2006)
6. Herbert, S., Marculescu, D.: Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In: ISLPED Pub., pp. 38–43 (2007)
7. Dwokin, M.: Recommendation for Block Cipher Modes of Operation: The CCM Mode for Authentication and Confidentiality, NIST Pub. 800-38C (2002)
8. POSIX Threads Programming (2006), http://www.llnl.gov/computing/tutorials/pthreads

Effective Object Identification through RFID Reader Power Control*

Shung Han Cho[1], Sangjin Hong[2], and Nammee Moon[3]

[1] System LSI Business, Samsung Electronics, South Korea
[2] Department of Electrical and Computer Engineering, Stony Brook University, USA
[3] Department of Mobile System, Hoseo University, South Korea

Abstract. In this paper, we present a RFID reader scheduling strategy for an effective identification by varying coverage through power control. A RFID reader is incorporated into a surveillance system which tracks objects with visual sensors. A required separation between objects is defined to avoid group identification with multiple objects. In order to reduce the power consumption of a RFID reader, the power cost is simply modeled with object positions to determine the activation time and the range of a RFID reader. The power cost model also considers the effect of added power consumption establishing group identification. A RFID reader is activated when the estimated power cost for the current sampling time is smaller than the estimated power cost for the next sampling time. The simulation results demonstrate that the proposed method reduces the power consumption with the effective object identification.

1 Introduction

In recent years, a green technology to save energy has received much attention in the research community. A surveillance system is one of the sensor based applications that require to minimize unnecessary power consumption of sensors. Some surveillance systems has visual sensors to track objects as well as identification sensors to identify tracked objects [1][2][3]. In order to collaboratively use two different types of information, an approximated coverage of a RFID reader is shared between visual sensors and a RFID reader as common reference information [3]. The system maintains the sets of estimated positions and identifications for each coverage of a RFID reader. Estimated positions are identified with their identifications if the number of added or subtracted elements in the sets is equal to each other [4]. In order to check the variations of the sets, the sensors are always activated to identify objects even though any objects may not exist in the surveillance region. However, it is not trivial to minimize the power consumption of a RFID reader. The activation time

* This research is supported by the International Collaborative R&D Program of the Ministry of Knowledge Economy (MKE), the Korean government, as a result of Development of Security Threat Control System with Multi-Sensor Integration and Image Analysis Project, 2010-TD-300802-002.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_26,

depends on unknown object trajectories and the activation range is affected by the separation issue between multiple objects. Therefore, an efficient scheduling strategy for a RFID reader is required to reduce the unnecessary power consumption.

In this paper, we present a RFID reader scheduling strategy by varying coverage through power control. The required separation between objects is defined to check the possibility of single identification. The simple power cost model is presented by considering object positions to determine the activation time and the range of a RFID reader. To maximize single identifications, the power cost model also considers the effect of added power consumption establishing group identifications. A RFID reader is activated when the estimated power cost for the current sampling time is smaller than the estimated power cost for the next sampling time. We compare the proposed power control method with the minimum power consumption method and the maximum power consumption method in terms of the identification performance and the power consumption.

2 Object Identification through Power Control

2.1 Known Object Trajectory

A RFID reader normally remains sleep state to reduce the power consumption when there are no object activities for the identification. The RFID reader is activated when an object and a RFID reader are at the minimum distance from each other. d_i^k denotes the Euclidean distance between RFID reader R^k and object O_i within $R_{\max}$. When a RFID reader is activated to identify multiple objects as shown in Fig. 1, the success of a single identification depends on the tangential distance between objects. $d_{i,j}^k$ denotes the difference between the tangential distances between d_i^k and d_j^k to the center of RFID reader R^k. If $d_{i,j}^k$ is smaller than Δd, a group identification can be established for objects O_i and O_j. Δd denotes the required separation of the tangential distance between objects to achieve a single identification. To avoid group identification, a RFID reader is not activated event at the minimum distance when the distance between objects is smaller than or equal to Δd.

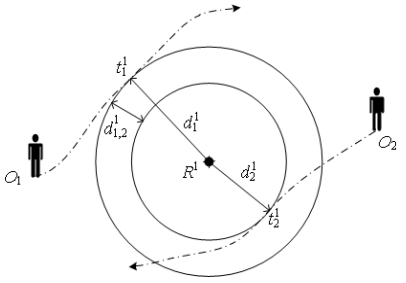


Fig. 1. Basic operation with a single RFID reader to identify multiple objects

When multiple RFID readers are installed in the environment, each RFID reader is assigned to identify each object. The Euclidean distances for the pairs of RFID readers and objects are calculated, and each pair of a RFID reader and an object is determined by the minimum distance. If the determined pair is violated by Δd, the second minimum distance is used. The RFID reader scheduling minimizes the power consumption but there is a possibility to delay the identification because of the required separation.

2.2 Unknown Object Trajectory

In reality, it is difficult to accurately calculate the time instant of the minimum distance because object trajectories are not perfectly known to the system. Hence, the system tries to activate a RFID reader whenever an object enters the coverage. Although the immediate activation of a RFID reader shortens time to identify an object, it increases the power consumption with the maximum coverage.

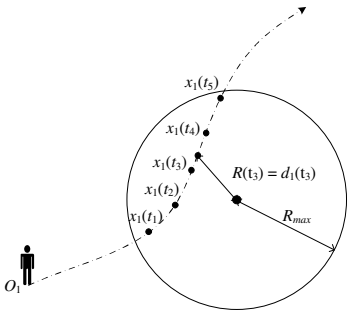


Fig. 2. Identification of an object when an object trajectory is not perfectly known

In order to reduce the power consumption, the activation time of a RFID reader is determined by estimating power consumption. The system estimates the power consumption with the distance between an object and a RFID reader every sampling time as shown in Fig. 2. The power consumption of a RFID reader is usually proportional to its activated range. $d_i(t)$ denotes the distance between object O_i and a RFID reader at time *t*. The power consumption for the next sampling time is calculated with the estimated position of an object with the linear trajectory model during the sampling period of visual sensors. $\hat{d}_i(t)$ denotes the distance to the estimated position at time *t*. By comparing $d_i(t)$ with $\hat{d}_i(t+1)$, the system determines the activation time and the range of a RFID reader. If $\hat{d}_i(t+1)$ is smaller than $d_i(t)$, the system waits for the next sampling time to use less power with the range of $\hat{d}_i(t+1)$. Otherwise, the system activates a RFID reader with the range of $d_i(t)$ because the smaller range is not guaranteed later.

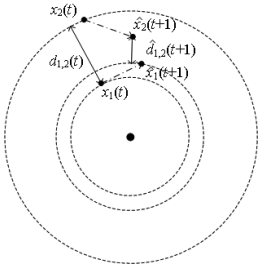


Fig. 3. The delay of single object identifications by the power cost estimation using only object positions

Although the scheduling of a RFID reader by searching all the cases may reduce the power consumption, the number of cases increases exponentially to the number of objects within the coverage of a RFID reader. Also, if other objects enter the coverage of a RFID reader at the next sampling times, the anticipated identification with the determined scheduling of a RFID reader may not occur. In order to efficiently determine the scheduling of a RFID reader for multiple objects, the power cost is simply estimated with the distance of the outermost position to the center of a RFID reader and is represented by

$$\hat{P}(t) = f(\max_i d_i(t)), \tag{1}$$

where $\hat{P}(t)$ denotes the estimated power cost at time *t*. When the system determines the distance of the outermost object, the system considers a possibility of establishing single identification disjointed from group identification. If the outermost position has a chance to establish single identification by the identification of other positions, the corresponding outermost position is excluded in (1). The system activates a RFID reader at the current sampling time only if $\hat{P}(t)$ is less than $\hat{P}(t+1)$. However, objects may move out of the coverage of a RFID reader at the next sampling time even though $\hat{P}(t)$ is greater than $\hat{P}(t+1)$. If any estimated distances $\hat{d}_i(t+1)$ are greater than $R_{\max}$, the system activates a RFID reader at the current sampling time to identify objects as much as possible. Once it is determined to activate a RFID reader with its outermost position, the system also determines the order of the range according to the increasing order of $d_i(t)$ to the outermost position. If multiple distances are in relation to group identification within Δd, the farthest distance among them is selected.

However, the power cost estimation with only the distance of the outermost position may delay single identification. For example, when two objects are in the coverage of a RFID reader as shown in Fig. 3, $\max_i d_i(t)$ is greater than $\max_i \hat{d}_i(t+1)$ but $d_{1,2}(t)$ is greater than Δd and $\hat{d}_{1,2}(t+1)$ is smaller than

Δd . According to the power estimation with the outermost position, the system waits for the next sampling time. Then, at the next sampling time, the system activates a RFID reader with the range of $\max_i d_i(t+1)$ but establishes group identification.

In order to avoid delaying single identification, the power cost model considers the effect of group identification by

$$\hat{P}(t) = \alpha_n \cdot f(\max_i d_i(t)), \tag{2}$$

where α_n denotes the weight factor for establishing identification with n positions with the outermost position. $n = 1$ indicates single identification and $n > 1$ indicates group identification for n objects. $\alpha_n (n > 1)$ is set to be larger than α_1 because group identification is required to activate a RFID reader later to establish single identification.

3 Simulation

Fig. 4 shows the comparison simulation with three methods to identify three objects with a single RFID reader. The first method is the minimum power consumption method to schedule of a RFID reader under the assumption that object trajectories are completely known to the system. The second method is the maximum power consumption method to activate a RFID reader as soon as possible for entering

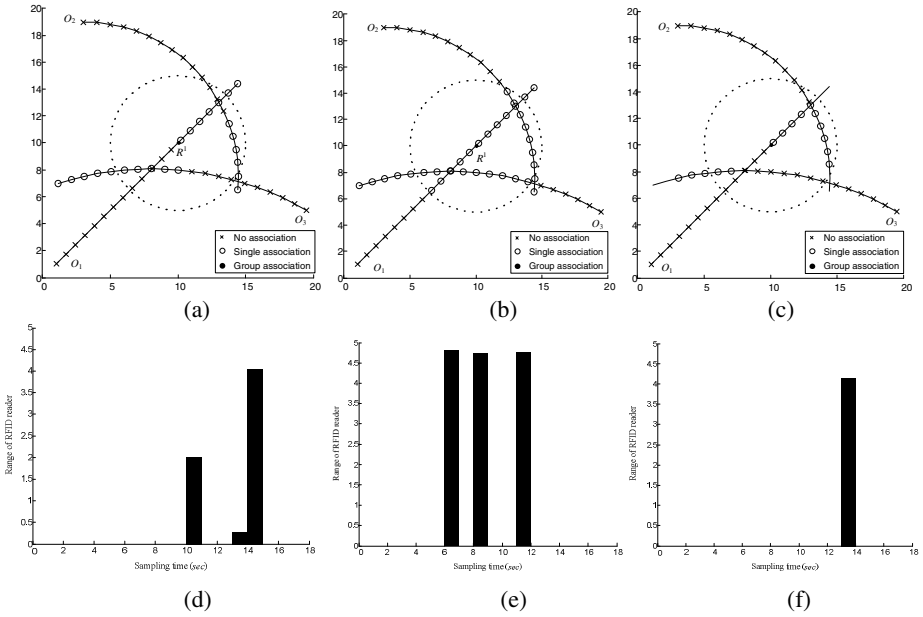


Fig. 4. Comparison simulation of a single RFID reader to identify three objects

objects under the assumption that object trajectories are unknown. The third method is the proposed power control by using the weighted power cost model. RFID reader R^1 is placed at $(10m,\ 10m)$ with the maximum range $5m$. The sampling period of object trajectories is set to be $0.5sec$ and Δd is set to be $0.2m$. α_n is simply set to the number of identified positions as a group and one for single identification. The initial positions of objects are $(1m,\ 1m)$, $(3m,\ 19m)$ and $(19.5m,\ 5m)$ respectively. Fig. 4 (a), Fig. 4 (b) and Fig. 4 (c) show that objects are identified at different sampling times by the three different methods respectively. When object trajectories are known, the RFID reader is activated when objects are at the minimum distance to the RFID reader. Although it uses the minimum power, it takes longer time to identify objects as shown in Fig. 4(a) and Fig. 4(d). On the other hand, when object trajectories are unknown, the RFID reader is activated whenever it has a chance to identify objects. It shortens time to identify objects but increases the power consumption because of the long range activation as shown in Fig. 4(b) and Fig. 4(e). Fig. 4(c) and Fig. 4(f) demonstrates that objects are identified faster than the first method with the less power consumption than the second method.

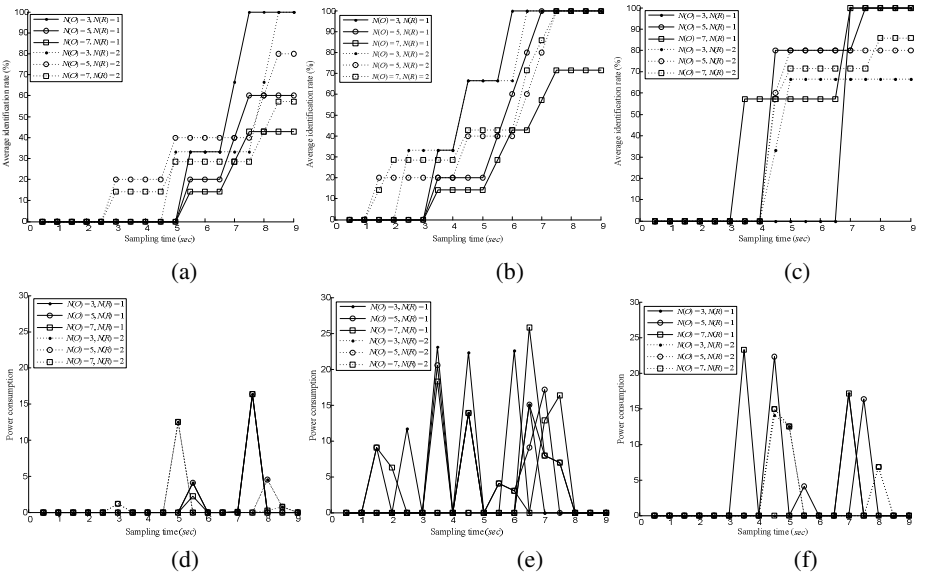


Fig. 5. The comparison simulations as varying the number of RFID readers and objects $N(O)$ denotes the number of objects and $N(R)$ denotes the number of RFID readers

Fig. 5 shows the simulation results to compare the performance of the three methods as varying the number of RFID readers and objects. R^1 and R^2 are placed at $(5m,\ 10m)$ and $(15m,\ 10m)$ with the maximum range $4m$. The initial positions of other objects O_4, O_5, O_6 and O_7 are $(19.5m,\ 15m)$, $(20m,\ 10m)$,

$(18m,\ 2m)$ and $(20m,\ 8.5m)$ respectively. The average identification rate includes group identification as well as single identification, and the power consumption is calculated by simply summing the squared range of a RFID reader. With the first and second method, as the number of objects increases, it takes longer time to identify objects because it is difficult to satisfy Δd . The average identification rate with the first method is more affected by the number of objects because it needs to satisfy both Δd and the minimum distance condition. On the other hand, with the second method, objects are identified much faster but the power consumption is much higher than the first method. The proposed power control shows the better identification performance than the first method and the less power consumption than the second method.

4 Conclusion

In this paper, we present the RFID reader scheduling method for the effective object identification by varying coverage through power control. The required separation between the object positions is defined to check the possibility of single identification. The simple power cost model is presented with the outermost position of objects and is also weighted by the expected group identification. A RFID reader is activated when the estimated power cost for the current sampling time is smaller than the estimated power cost for the next sampling time. The simulation results show that the proposed method reduces the power consumption with the effective object identification.

References

1. Schulz, D., Fox, D., Hightower, J.: People tracking with anonymous and ID-sensors using Rao-Blackwellised particle filters. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (August 2003)
2. Shin, J., Kumar, R., Mohapatra, D., Ramachandran, U., Ammar, M.H.: ASAP: A camera sensor network for situation awareness. In: Tovar, E., Tsigas, P., Fouchal, H. (eds.) OPODIS 2007. LNCS, vol. 4878, pp. 31–47. Springer, Heidelberg (2007)
3. Cho, S.H., Hong, S., Moon, N., Park, P., Oh, S.-J.: Object association and identification in heterogeneous sensors environment. EURASIP Journal on Advances in Signal Processing 2010, Article ID 591582 (November 2010)
4. Cho, S.H., Nam, Y., Hong, S.: Association and identification in heterogeneous sensors environment with coverage uncertainty. In: Proc. of IEEE Int'l Conf. on Advanced Video and Signal Based Surveillance (September 2009)

Market-Based Resource Allocation for Energy-Efficient Execution of Multiple Concurrent Applications in Wireless Sensor Networks

Mo Haghighi

Department of Computer Science, University of Bristol, Bristol, UK
Mo.Haghighi@Bristol.ac.uk

Abstract. Many engineering disciplines have become reliant on WSNs in order to detect and track certain events of interest by monitoring various variables, through a number of specially distributed wireless sensors. Due to resource constraints of sensor hardware, traditional WSN applications involved exchanging an excessive amount of data, usually in an offline mode, between sensor nodes and a central unit, in order to apply computational analysis on the captured data. New sensor devices however, are equipped with more powerful resources and capable of running multiple concurrent processing, and applying computational data analysis can be implemented online and often in a distributed fashion. In this paper we will investigate the application of market-based algorithms for energy management, tasks allocation and resource coordination in WSNs with multiple concurrent applications. We will also propose a number of algorithms for calculating costs and utilities for multi-paradigm application requirements.

Keywords: Market-based, Auction-based, WSN, Sensor, Utility, Sensomax, Concurrency.

1 Introduction

Wireless Sensor Networks have gained increasing attention in many fields including monitoring buildings and other structures, healthcare, military, and recording environmental conditions. Every WSN consists of a number of sensor devices (nodes) with wireless communication capabilities, which are distributed across an environment, to detect and track changes in the variables of interest. The node devices suffer from a number of hardware constraints such as limited energy sources, processor-power and memory-space. WSNs often need to operate unsupervised for long periods of time in remote environments where physical access is very limited or impossible. Every sensor node comes integrated with a software component that regulates access to on-board hardware resources and manages communications with other network devices whist maintaining a reasonable energy level in order to prolong network longevity. This software component is often referred to as Middleware. In centralized networks, the administrative role of the middleware is very minimal because it only

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_27,

takes charge of the inner-node processing and communicating the results/commands with a central unit. In distributed networks however, other nodes' current states and properties, as well as the overall network energy requirements, all need to be taken into account before making any decision. Traditional sensor networks mostly involved serialized actions, which needed to be executed sequentially. On one hand, that was due to resource-constraints of WSNs hardware in general; and on the other hand, WSN applications were not advanced enough and therefore required less sophisticated computational data processing. *By parallelizing computational tasks across a WSN, problems associated with power efficiency, data loss, and finite communication ranges will be minimized while providing a framework for the autonomous, in-network processing of sensor data* [6]. There are various methods to facilitate autonomous and efficient parallelization in WSNs including approaches based on game theory, statistical analysis, and market-based techniques.

Market-based techniques attempt to envision the network as an auction where a computational task is offered as a product for sale by the end-user. A number of sensor nodes, modeled as buyers, state their interest for the offered product by declaring their bidding price to the seller. The bidding price given by the buyers are calculated in proportion to a number of factors such as the amount of energy, processing and memory required to execute the task. Each node, based on its currently available resources may calculate a different cost for executing a task. The nature of cost estimation is very dynamic as every node may offer a different price for the same task at different instances, since the state of every node is constantly changing according to its remaining energy and its running tasks. The distributed nature of market-based techniques makes them suitable to induce localized in-network, in-cluster and in-node autonomies in WSNs; and the lack of centralized control means there is not a single point of failure.

Sensomax is an agent-based middleware, which has been developed in Java and is capable of running multiple concurrent applications in a decentralized and multi-paradigm mechanism. Sensomax was originally developed in Java ME to run on a network of Sun Spot devices [10]. However it was later converted into Java SE and SE-embedded to interoperate on various Java-enabled devices including PCs and Raspberry Pi [11]. As an object-oriented WSN middleware, Sensomax features a number of unique functionalities, as a result of exploiting core Java APIs including threading, sockets and garbage collection, which ease the process of running soft computational algorithms without imposing heavy overheads on the resources.

In this paper, we will investigate the feasibility and the effectiveness of applying market-based techniques in Sensomax's task distribution, in order to simplify the aforementioned complexities and implement an autonomous resource-allocation mechanism in which the following objectives can be achieved:

- Maximizing the inner-node collective utility of multiple concurrent applications' processing, whilst minimizing the cost of each individual application execution;
- Maximizing the overall hardware resources' utilities, whilst minimizing the cost of accessing each individual hardware resource.

2 Architecture and Cost Equations

Agents are the main communication means in Sensomax's architecture; therefore most task-related operations are highly dependent on how fast the agents are processed. A task is initially deployed as an agent onto the network and decomposed into multiple subtask agents inside a node. Each subtask agent is further refined and processed by different components based on its properties. We demonstrated the entire operation in [1] and [2]. In our market-based model, each task corresponds to a specific time interval, which is equal to the total duration of time for that task agent to be processed by a node. As we explained before, each node may process an agent faster or slower based on the number of on-going processing. Hence, each node advertises a different price for performing the same task.

The other major factor for calculating the cost is the nature of the task, which is highly dependent on its operational paradigm and defined by its requirement Sensomax can inter-operate in four different paradigms: Event-Driven; Time-Driven; Data-Driven; and Query-Driven. Subtasks are separated into one of these categories and executed in their own paradigm-driven execution space. Based on our experiments each paradigm takes a variable amount of time, and the list of paradigms in the previous sentence is ordered by time-costs, with the event-driven taking the longest processing time and the query-driven the shortest. In general, the shorter the agent takes to be processed, the less energy will be consumed (since the node can switch into sleep mode or process other queuing agents), which ultimately results in longer lifetime of the node and the network.

There are a number of factors involved in the processing time of the agents according to their paradigms. That is, executing event-driven agents involves setting up a number of components to closely monitor a variable for an event of interest, whereas the time-driven agents only requires a timer component to regularly check up on a variable; and finally query-driven agents are instant raw queries received into the node, demanding an immediate measurement of a specific variable.

Having done a thorough study of the existing research on WSN market-based resource allocations in [3-7] and [9], a number of algorithms have been derived that are exclusively tied to multi-paradigm execution of agents in Sensomax. Each paradigm owns a different cost function (equation) for calculating the cost associated with the execution of its own type of agent based on its remaining capacity. Every function is also dependent on the number of agents in the buffer and the number of parallel tasks.

Equation 1 and 2 calculate the overall costs associated with query-driven and data-driven tasks, which have derived from query agent α and data agent β respectively.

$$P_Q(\alpha) = \sum_{i=1}^{i=S^{C_Q}} (\tau_{\alpha_i}.\partial_Q) \quad | \ \alpha_i \in C_Q,\ \partial_Q = \lambda_Q,\ \lambda_Q = \left(\frac{N_B}{N_R}\right).\varepsilon_Q \tag{1}$$

$$P_D(\beta) = \sum_{j=1}^{j=S^{C_D}} \sum_{i=1}^{i=S^{C_Q}} [P_Q(\alpha_i) + (\tau_{\beta_j}.\partial_D)\,] \quad | \ \beta_j \in C_D, \alpha_i \in C_Q, \partial_D = \lambda_D - \lambda_Q, \lambda_D = \left(\frac{N_D}{N_R}\right).\varepsilon_D \tag{2}$$

Equation 3 and 4, on the other hand, calculate the overall costs associated with time-driven and event-driven tasks, which have been derived from time agent γ and event agent δ respectively.

$$P_T(\gamma) = \frac{\sum_{j=1}^{j=S^{C_T}} \sum_{i=1}^{i=S^{C_Q}} K_i \,.\left[P_Q(\alpha_i)\,.\,\left(\frac{L}{1000}\right)\right] + (\tau_{\gamma_j}.\partial_T)^{K_i}}{\Delta T/_M} \mid \gamma_j \in C_T, \alpha_i \in C_Q, \partial_T = \lambda_T - \lambda_Q, \lambda_T = \left(\frac{N_T}{N_R}\right).\varepsilon_T \quad (3)$$

$$P_E(\delta) = \frac{\sum_{j=1}^{j=S^{C_E}} \sum_{i=1}^{i=S^{C_T}} \sum_{h=1}^{h=G} P_T(\gamma_j) + \left(\tau_{\delta_i}.\partial_E\right)^{F_h}}{\Delta T/_M} \mid \gamma_j \in C_E,\ \delta_j \in C_T,\ \partial_E = \lambda_E - \lambda_E, \lambda_T = \left(\frac{N_E}{N_R}\right).\varepsilon_E \quad (4)$$

As Table 1 outlines, C_x stands for the whole collection of subtask agents integrated in the agent type x, with τ_x as the total time interval and ∂_x as the total delay caused by processing agent x, based on the ratio of N_B: the total number of agents queuing in the buffer to N_R: number of agents already in the process, multiply by the constant ε_x, which denotes the total processing delay exclusively associated with each type of agent in Sensomax.

Table 1. Outlines

Description\Type	Agent-Driven	Data-Driven	Time-Driven	Event-Driven
Agent	α	β	γ	δ
Time Interval	τ_{α_i}	τ_{β_j}	τ_{γ_j}	τ_{δ_i}
System Constant	ε_Q	ε_D	ε_T	ε_E
Processing Delay	∂_Q	∂_D	∂_T	∂_E
Subtasks Collection	C_Q	C_D	C_T	C_E

In all equations throughout this paper, S_x denotes the size of collection C_x, ΔT stands for the total lifetime of the application and M is the remaining lifetime of the node in milliseconds. In Equation 3, L stands for the time interval associated with each time-driven subtask (milliseconds) and K stands for the number of variables that need to be monitored simultaneously in every sub-task. In Equation 4, F stands for the number of dependent events in each individual subtask with their total number equating to G.

As was pointed out earlier, query-driven agents take the lowest overhead since they get executed instantly; therefore nodes' and applications' lifetimes are not taken into account for these two types of agents. One important point to notice here is, data-driven and time-driven agents potentially contain a number of query agents plus some extra processing for retrieving data from the resources and taking measurements respectively; that is why equations 2 and 3 integrate equation 1 as well. Also equation 4 takes account of all time-driven subtask agents that are embedded in the event-driven

ones, that is due to the fact that every event-driven agent contains a number of time-driven agents that are required to set up a number of timers for regularly checking up on variables of interest

3 Case Study

We conducted a number of experiments to validate the effectiveness of market-based application in Sensomax's architecture, in order to achieve the objectives mentioned in Section 1. All experiments have been conducted in the SXCS [8] simulator on 1000 virtual nodes with the proposed cost equations integrated in them. The first phase of this experiment investigates the increase in the cost of executing different types of subtasks based on the number of nodes deployed in the network, according to the categories mentioned in the previous section.

The second phase of the experiment evaluates the impact of using market-based techniques on the performance of the nodes in large-scale scenarios.

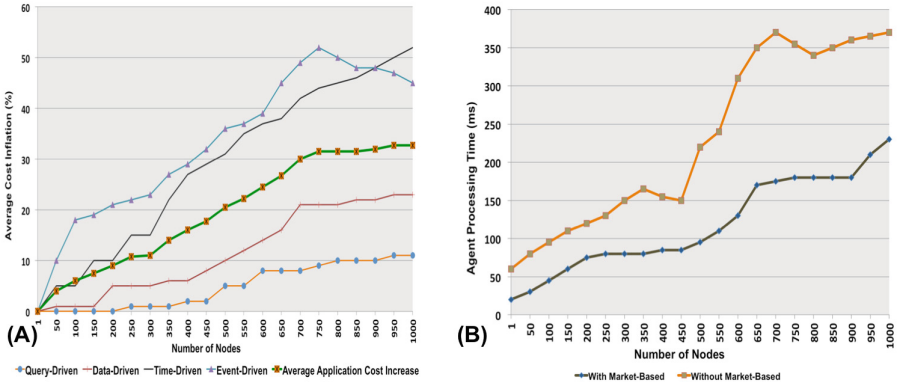


Fig. 1. Cost Inflation and Agent Processing Time vs. Network Density

Figure 1(A) shows the average cost inflation (in percentage) for all types of applications' subtasks according to the network density. Based on this figure, event-driven tasks (blue line) own the highest inflation and the query-driven subtasks (orange line) own the lowest. Green line shows the average application cost increase based on the average increases of all four types of tasks as the network density grows. Based on this figure, except time-driven tasks, all other paradigms' inflations gradually slow down after 700 nodes, which is quite significant in the case of event-driven tasks. The linear growth of time-driven tasks can be distributed to the number of frequent timers that regularly trigger the node and increase the overall processing.

Figure 3(B) on the other hand illustrates the average processing time of the agents in milliseconds with (blue line) and without (orange line) market-based techniques in Sensomax, which validates the efficient usage of such a technique and its impact on the overall performance of the middleware. According to this figure, such techniques in a competitive manner can significantly optimize the system's performance as the network scales up to incorporate a larger number of nodes in the form of auctioneers.

4 Conclusion

In this paper we have evaluated the effectiveness of market-based techniques in WSNs with multi-paradigm requirements. We have also proposed a number of equations for calculating the costs associated with executing such requirements. Utilizing market-based techniques proved to decrease the agent processing time, which results in longer periods that sensor nodes stay in the low-processing state (sleep mode), and ultimately prolongs the network longevity. Lowering the agent processing latency also proved to improve the applications' responsiveness and create more efficient execution environments for multiple concurrent applications in Sensomax. In our future research, market-based techniques will be exploited to optimize clustering and regulate resource allocations on the cluster level.

References

1. Haghighi, M., Cliff, D.: Sensomax: An Agent-Based Middleware For Decentralized Dynamic Data-Gathering in Wireless Sensor Networks. In: The 2013 International Conference on Collaboration Technologies and Systems, CTS 2013, San Diego, USA (May 2013)
2. Haghighi, M., Cliff, D.: Multi-Agent Support for Multiple Concurrent Applications and Dynamic Data-Gathering in Wireless Sensor Networks. In: The Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2013, Taiwan (July 2013)
3. Mullen, T., Avasarala, V., Hall, D.L.: Customer-Driven Sensor Management. IEEE Journal of Intelligent Systems 21, 41–49 (2006)
4. Edalat, N., Wendong, X., Chen-Khong, T., Keikha, E., Lee-Ling, O.: A price-based adaptive task allocation for Wireless Sensor Network. In: IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, pp. 888–893 (2009)
5. Elmogy, A.M., Khamis, A.M., Karray, F.O.: Dynamic complex task allocation in multisensor surveillance systems. In: 3rd International Conference on Signals, Circuits and Systems (SCS), pp. 1–6 (2009)
6. Zimmerman, A.T., Lynch, J.P., Ferrese, F.T.: Market-based computational task assignment within autonomous wireless sensor networks. In: IEEE International Conference on Electro/Information Technology 2009, pp. 23–28 (2009)
7. Cheng, W., Shengbin, L., Wei, L., Zongki, Y., Kanru, X.: A Price-Based Distributed Algorithm for Optimal Utility-Energy Trade-Off in Wireless Sensor Networks. In: 66th IEEE Vehicular Technology Conference, pp. 143–147 (2007)
8. Haghighi, M.: An Agent-based Multi-model Tool for Simulating Multiple Concurrent Applications in WSNs. In: Journal of Advances in Computer Networks (JACN), 5th International Conference on Communication Software and Networks, Malaysia (June 2013)
9. Khan, M.I., Rinner, B., Regazzoni, C.S.: Resource coordination in Wireless Sensor Networks by combinatorial auction based method. In: 3rd IEEE International Conference on Networked Embedded Systems for Every Application, pp.1–6 (2012)
10. Oracle, Sun Spot Programmer's manual, Release v6.0, Sun Labs, Oracle (2010)
11. Upton, E., Halfacree, G.: Raspberry Pi. USER GUIDE IS. John Wiley and Sons (2012)

Clustering Objects in Heterogeneous Information Network Using Fuzzy C-Mean

Muhammad Shoaib and Wang-Cheol Song*

Department of Computer Engineering,
Jeju National University, Jeju, Republic of Korea
muhammad.shoaib@live.com, phillo@jejunu.ac.kr

Abstract. In this paper we have proposed a fuzzy c-mean based clustering algorithm for categorization of different types of objects present in a heterogeneous information network. We have addressed a particular scenario in this paper when exact structure of objects and their relationships with other objects is either hidden or not known. We have performed the experiments on an agriculture information network and our results depicts that combining automatic extraction of structure of an information network with information objects can improve the quality of clustering.

1 Introduction

Information networks [4] are compound of different objects belonging to various types connected with each other with different relationships. An information network is very simple when its objects belong to same type. These types of information networks are called homogeneous information networks. On the other hand an information network can be so complex that its objects are connected with objects belong to more than one domain. These types of information networks are known as heterogeneous information networks [4, 8].

Information networks can be analyzed at two different levels, structure level and data level [4]. In structure level objects are analyzed based on their relationship with other objects, where at data level objects are analyzed based on specific values for relations [9, 4, and 6]. However, analyzing them separately does not generate as much useful results as their combination can produce [9]. In this work we have proposed an algorithm for analyzing objects in heterogeneous information networks by combining structure level and object level clustering techniques. We have used fuzzy c-mean algorithm for automatic extraction of relationships from multi-dimensional datasets. The proposed technique is based on following two steps.

We first cluster the objects based on their types that are identified using relationship that these objects have with other objects in a heterogeneous information network. In each cluster objects with the similar relations are organized. With this process we place the objects of similar type in one cluster. In the second step we

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_28,

apply the clustering technique on each cluster obtained in result of step 1. However this time we cluster the objects based on the values for the relationships they have. The benefits of our proposed approach are twofold. Firstly Our proposed method is more useful when the information about structure of network is either hidden or cannot be analyzed manually. Secondly the relations that exist between objects still remain same after the clustering algorithm is applied as we do not convert heterogeneous network into homogeneous network but we treat it as a heterogeneous network at all stages presented in entire process. We used fuzzy membership function in order to create soft and overlapping clusters so that objects with multiple relationships can be clustered in multiple clusters.

The rest of the paper is organized as follows; section 2 explains proposed algorithm and its mathematical model followed by experiments and results in section 3. Finally section 4 concludes the paper.

2 Clustering Heterogeneous Information Network

Here we first have defined the concepts related to heterogeneous information network in order to formalize the concept of clustering.

Definition 1. *Information Network: Given a set of atomic types* $T = \{t_1, t_2, t_3, t_4 \dots t_n\}$, *set of objects* $= \{O_{t=1}^{T}\}$ *where* O_i *is set of objects belong to type* t_i *and set of tions* $\mathcal{R} = \{r_1, r_2, r_3, r_4 \dots r_n\}$, *a Description Graph* $G = (V, E)$ *is called an* ***information network*** *for if* $V \in$ *and* E *is a semantic relation on V and* $E \in \{V \times \mathcal{R} \times V\} \cup \{V \times \mathcal{R} \times I\}$ *where I belongs to class of literal values i.e. data type values.* ■

Let $G = (V, E)$ a simple information network, we defined $G = (V, E, W)$ as a ***weighted information graph*** such as $E = \{ e \in E \wedge e = (u, r, v, w) : u, v \in V \}$. Weight is used in an edge to define the importance and strength of relationship among two objects and is a bi-side property. Using weight we analyze the strength of relationship that exists between two different objects i.e. how much these two objects are necessary for each other.

Definition 2: *Similar Objects: Given objects* O_i *and* O_j*connected with set of objects* $\{(R_i, U_i)\}$ *and* $\{(R_j, U_j)\}$ *respectively, where* $R_i, R_j \subseteq \mathcal{R}$ *and* $U_i, U_j \subseteq$, O_i *and* O_j *are said to be similar objects if and only if there exists direct mapping* $U_i \rightarrow U_j$ *and* $\left(\forall\, x \in U_i, y \in U_j \,\exists\, T(x) = T(y)\right) \wedge \left(\forall\, a \in R_i, b \in R_j \,\exists\, a = b\right)$

Definition 3: *Equal Objects: Given objects* O_i *and* O_j*connected with set of objects* $\{(R_i, U_i)\}$ *and* $\{(R_j, U_j)\}$ *respectively, where* $R_i, R_j \subseteq \mathcal{R}$ *and* $U_i, U_j \subseteq$, O_i *and* O_j *are said to be equal objects if and only if there exists direct mapping* $U_i \rightarrow U_j$*and* $\left(\forall\, x \in U_i, y \in U_j \,\exists\, x = y\right) \wedge \left(\forall\, a \in R_i, b \in R_j \,\exists\, a = b\right)$

2.1 Finding Schema Level Similarity

Given O_i *and* O_j *are two different objects, where* O_i *has connected nodes* $X = \{N_{out}(O_i)\}$ *and* O_j *has values for a set of attributes* $Y = \{N_{out}(O_j)\}$. *We first define*

set of types for $X_t = \{\forall x \in X \mid x = (r, T(v))\}$ *and* $Y_t = \{\forall y \in Y \mid y = (r, T(v))\}$.

Now we are ready to define the Schema Level Similarity of objects in term of set operations. For $x \in X_t$ *and* $y \in X_t$ *we call them equal if and only if* $(r(x) = r(y)) \wedge (T(v(x)) = T(v(y)))$ *Equation 1 explains the mathematical model for* $SSim$

$$SSim\ (O_i, O_j) = \frac{2(|X_t\ \cap Y_t|)}{|X_t| + |Y_t|} \quad \text{EQ(1)}$$

When network is weighted information network $SSim$can be found using equation 2 formula

$$SSim\ (O_i, O_j) = \frac{\sum_{e \in (X_t \cap Y_t)} w(e)}{\sum_{e \in X_t} w(e) + \sum_{e \in Y_t} w(e)} \quad \text{EQ(2)}$$

2.2 Object Level Similarity

In order to find the similarity between two objects irrespective of object type – for example finding two patients of having two different diseases – we need to take the original values for properties of that object.

Let $O_i = \{(r_i, x_i, w_i)\}$ *and* $O_j = \{(r_i, x_i, w_i)\}$ *where* $r_i, r_j \in \mathcal{R}$, $x_i, x_j \in O \cup I$ *and* $w_i, w_j \in \mathbb{R}$. The Object Similarity *(*$OSim$*)* can be defined as equation 3

$$OSim\ (O_i, O_j) = = \sum_{i=1}^{n} |x_i w_i - x_j w_j| \quad \text{EQ(3)}$$

2.3 Construction of Clusters

Definition 4: *Information Network Cluster: Given a sub graph* $C(V^*, E^*) \subseteq G(V, E, W)$ *where* $V^* \subseteq V(G)$ *and* $E^* = \{e \in (E(G) \mid e = (u, r, v)\ with\ u, v \in V^*\ and\ r \in \mathcal{R}\}$ *can be said an* ***information network cluster*** *if and only if there exists no object* $O_i, O_j \in V^* \wedge Sim(O_i, O_j) < \tau$ *where* $\tau \in \mathbb{R}$ ■

We discuss two types of clustering for heterogeneous information network based on fuzzy c-mean clustering algorithm. 1) Autonomous clustering in which there are no defined center points. This can also be called un-supervised clustering. 2) Manual clustering in which we first define the disjoint objects by our self as cluster centroids [2, 7]. In autonomous clustering the algorithm first selects centroids randomly and then evolves these centroids for each cluster iteratively in rest iterations. As the results of this iterative process best cluster centroids are chosen for each cluster. A cluster centroid is called best centroids if it has maximum common relationships or common values with respect to other objects that are present in the same cluster.

2.4 Fuzzy C-Mean Algorithm

In fuzzy c-mean clustering algorithm [1] each object is made part of some cluster based on membership function. Distence for each object from center of each cluster is measured. We already have defined the measurement of difference between two

objects therefore here will just briefly introduce the fuzzy c-mean membership function. The main function for fuzzy c-mean that is needed to be minimized has been explained in equation 4.

$$J(O_i, V_j) = \sum_{j=1}^{N} \sum_{i=1}^{c} \mu_{ij} \; Diff(O_i, V_j) \quad - \quad \text{EQ(4)}$$

$$Diff(O_i, V_j) = 1 - \left| \frac{1}{\left(SSim(O_i, V_j)\right) + \left(OSim(O_i, V_j)\right)} \right| \quad \text{EQ(5)}$$

here V_j is the cluster centroid of the cluster *j*. that can be computed using equation 8 and equation 7.

$$\mu_{ij} = 1 \Big/ \sum_{k=1}^{c} \frac{Diff(O_i, V_j)}{Diff(O_i, V_k)} \quad \text{EQ(6)}$$

$$V_j = \frac{\sum_j^N \left(\mu_{ij} \left(\sum\{(x, w)\}_j\right)\right)}{\sum_j^N \mu_{ij}} \; O_j = \{(r_i, x_i, w_i)\} \quad \text{EQ(7)}$$

3 Experiments and Results

In this section we apply our proposed clustering technique on agriculture information network, a heterogeneous information network connecting 5 different object types among each other. Firstly, we discuss how the clustering algorithm behave while creation of clusters of objects in order to find different types of objects present in the information network. Table 1 presents the number of classification and misclassification. We obtained good results as all more than 90% of the objects belonging to all classes were classified in their correct object type. In order to cross check the classification of the objects in different clusters we added a hidden object type in the dataset for each object. Once the clustering algorithm created all clusters using the step 1 in section 3.3 we used the hidden labels to match the accuracy of the clustered.

Table 1. Objects of Agriculture information network

Type	Nos	Attributes
Crop	20000	Name, Size, Color, Family, season
Soil	2000	Name, Family, Organisms, Texture
Fertilizer	1000	Name, Family, Soil-acidification
Herb	50000	Name, Family, Type, Color
Pests	1000	Name, Family, Season, Control-Method

The threshold value for assigning the membership was kept 75%. Two major reasons were identified for misclassifications have been noticed. First reason was missing relation types, for example all crops have relationship with soil, herbs, fertilizer, and pests. If two are more than two relationships were missed the membership function failed to assign the membership for cluster of crops' objects.

Figure 1 presents the accuracy of classification framework with respect to threshold value that was set for soft classification. Because the objects in our experimental dataset belonged to different object types therefore clustering without schema was not as much useful as clustering with schema was observed.

Next, Figure 2 presents the classification behavior with respect to no of cluster centroids. An increasing number of cluster centroids increase the performance and optimization of the clusters. When we have more cluster centroids this means more accurate clustering can be done because of having more membership functions. We chose 0.1 to 0.5 percent from the total objects as the cluster centroids randomly. However these cluster centroids was updated iteratively.

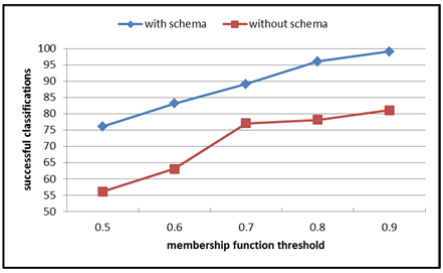


Fig. 1. Impact of threshold, membership function vs. accuracy of classification

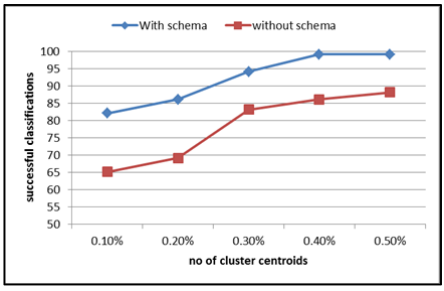


Fig. 2. Impact of cluster centroids

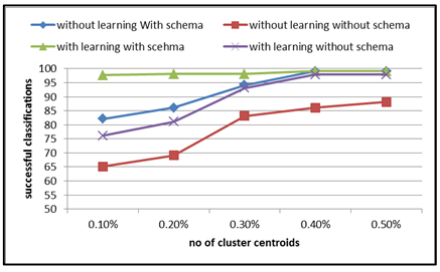


Fig. 3. Impact of Learning, cluster centroids vs. accuracy of classification

Learning has always played an important role in impeding the performance of computational systems. Figure 1, 2 and figure 3 presents cluster performance with and without learning. We also observe a linear relationship between learning ratio and classification output. As expected learning showed huge impact on the clustering process. Another interesting finding was relationship between number of centroids and learning. We found that in those scenarios when we can't define the cluster centroids because of any reason, increasing number of centroids can fulfill this limitation.

4 Conclusions

In this work we have presented a clustering algorithm based on fuzzy c-mean clustering algorithm in order to perform clustering on large scale unknown heterogeneous information networks. We have combined structure and object level clustering to improve quality of the clustering particularly when the objects are connected with each other and network structure is not available. We have performed experiments on an agriculture information network in order to study quality of the proposed algorithm. We discover, working with schema and objects together result as more precise clusters then creation of clusters separately.

References

1. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. 16(3), 645–678 (2005)
2. Wang, L., Leckie, C., Ramamohanarao, K., Bezdek, J.: Automatically determining the number of clusters in unlabeled data sets. IEEE Trans. Knowl. Data Eng. 21(3), 335–350 (2009)
3. Zheng, L., Li, T., Ding, C.: Hierarchical Ensemble Clustering. In: 2010 IEEE International Conference on Data Mining, pp. 1199–1204 (2001)
4. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge Univ. Press (2010)
5. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J.: SCAN: A structural clustering algorithm for networks. In: Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD 2007), San Jose, CA (August 2007)
6. Szabo, A., Castro, L.N., Delgado, M.R.: FaiNet: An Immune Algorithm for Fuzzy Clustering. In: Proc. WCCI 2012, IEEE World Congress on Computational Intelligence. IEEE press (2012)
7. Serban, G., Campan, A.: A New Core-Based Method For Hierarchical Incremental Clustering. In: Proc. the Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2005). IEEE (2005), doi:10.1109/SYNASC.2005.9
8. Wang, T., Srivatsa, M., Agrawal, D., Liu, L.: Modeling data flow in socio-information networks: a risk estimation approach. In: Proc. 16th ACM Symposium on Access Control Models and Technologies, pp. 113–122 (2011)
9. Han, J., Sun, Y., Yan, X., Yu, P.S.: Mining knowledge from databases: an information network analysis approach. In: Proc. SIGMOD 2010: 2010 ACM SIGMOD International Conference on Management of Data, pp. 1251–1252 (2010)

Better Induction Models for Classification of Forest Cover

Hyontai Sug

Division of Computer & Information Engineering, Dongseo University,
47 Jurye-ro, Sa-sang-gu, Busan 617-716, Korea
hyontai@yahoo.com

Abstract. The data set of forest cover types based on cartographic data consists of very large data set of 581,102 instances. So, decision tree-based data mining methods that need relatively less computing resources could be used for better classification models. Random forests consisting of multitude of special decision trees are known to be a good data mining tool, and a technique based on grid search of random forests was investigated to find very accurate classifier. Experiments showed that a classifier of high accuracy could be found for the data set of forest cover types.

Keywords: Data mining, forest cover type, decision trees, random forests.

1 Introduction

Many data mining specialists who are interested in data mining of very large data sets consider decision tree-based algorithms as good data mining tools because of their good scalability [1]. Decision tree algorithms build each branch of the tree based on some greedy split methods, so the tree construction time is relatively fast.

Lots of efforts have been devoted to build better decision trees [2, 3], and the algorithms have relatively long history of development and have been favored by many data mining experts [4]. Even though decision trees are very useful data mining tools, it is known that a weak point of decision trees is that decision trees sometimes show poor accuracy compared to other data mining methods like neural networks. Because decision tree algorithms prefer major classes to find the best accuracy and fragment training data sets, they neglect the classification of minor classes, so that decision trees show some poorer accuracy in many data sets than some other data mining algorithms like neural networks. Major classes are classes where classification is more accurate and have more instances in the classes [5]. Note that in neural networks all inputs of training instances are treated more equally, because the effect of training is dispersed in the networks as a form of weights [6].

Because decision tree induction algorithms fragment a training data set based on their branching criteria during tree building process, the training data set is drained relatively quickly, hence, the performance of trained decision trees is more heavily dependent on the training data sets. So, as a way to mitigate the problem BAGGING [7] was suggested for mostly small and medium-sized data sets. In BAGGING several

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_29,

equally sized training sets are made based on sampling with replacement, and trained knowledge models vote for classification or prediction. Random forests suggested by Breiman [8] are based on BAGGING, and use many special decision trees. The special decision trees are made by some random selection on attributes to generate the trees in the forests, and do no pruning. The performance of random forests is known to be as good as the performance of SVM [9] that is known to be one of the most accurate binary classifier [10].

For better classification models of the target data set, 'forest cover types, a few research results have been published due to the size of the data set after the first paper related to the task was published by Blackard and Dean [11]. They used ANN and achieved accuracy of 70.58% based on samples. Li [12] applied scalable decision tree system due to the size of the data set and achieved accuracy of 91.86%. Bo et al.[13] achieved the accuracy of 97.4% with their GS(greedy stagewise) SVM in which they took class 2 as one class and all the other as another class, because the data set consists of 7 classes, and SVM can treat 2 class problem only. Trebar and Steele [14] also transformed the data set to make it binary classification problem for their SVM, so the new 7 data sets have class 1 and all the other classes as another class, and class 2 and all the other classes as another class, and so on. They experimented on several sample sizes and the achieved best accuracies are 96.69%, 95.89%, 99.34%, 99.89%, 99.57%, 99.43%, 99.78% for data set 1 to data set 7 respectively. But, the results of SVMs cannot be used directly for classification task. Because each SVM could classify differently for a new instance, we need some ensemble method to classify the new instance.

2 Suggested Method

The target data set for our data mining has geographic information of four wilderness areas of the Rawah, Comanche Peak, Neota, and Cache la Poudre wilderness areas of the Roosevelt national forest in northern Colorado. The total number of independent attributes is 54, and one class attribute. The total number of observations in 30×30 m raster cells is 581,102. The data cover 129,213 acres of the area. For more details about the data set, you may refer to Blackard and Dean's [11]. The class distribution is highly unbalanced, so that we can say that class 1 and class 2 are major classes, which consist of 36.5% and 48.8% respectively, while class 3 ~ class 7 are minor classes that occupy only 14.8%. Therefore, we can expect that random forests could perform better than decision trees.

In random forests, there are two parameters, the number of trees, say nT, and the number of attributes to pick randomly to generate each subtree in a tree in the forests, say nA. The recommended default parameter values of nA and nT for most data sets are square root of A and hundreds respectively, where A is the number of attributes of the target data set [8]. The default values are based on data sets of conventional property like mostly data sets having moderate number of attributes and medium size. But, our target data set has some different properties. The size of the data set is very large and has several minor classes. So, we need to find best parameters for our target data set.

Generating random forests in all combination of nT and nA values may not be a good idea, because small changes in nT and nA will not generate much difference compared to computing time requirement. Therefore, we want to generate random forests based on the property of the target data set. As for the nA value for our data set, because we have very large data set, we should try several nA value increasingly from the default value. But small increase may not generate some different results, so we increase nA value as multiples of the default value as we generate three different numbers of trees in the forests.

As for the number of trees in our random forests, Breiman recommended hundreds of trees rather than thousands of trees, and moreover, because we expect that the bootstrap method [15] will not be as effective for our data set of very large size as for the cases of medium-sized data sets, we limit three sets of trees in the forests; 50, 100, and 150 trees. The following is rough procedure of the experiment:

```
Procedure:
  for nT := {50, 100, 150} do
    nA := 7 /* the square root of 54 */
    do while nA ≤ 54;
      Run random forests algorithm with nA and nT;
      if nA ≥ 49 then nA := 54
        else nA := nA + 7
       end if;
    end while;
  end for;
end.
```

A PC in window 7 operating system was used having AMD Phenom 2 processor with 20 GB of main memory and 2TB hard disk, and random forests in Weka [16] were utilized. Weka is a comprehensive data mining package developed with Java. 10-fold cross validation was used in the experiment for more objective experiment. Table 1 has the result of experiment. In the table, nT represents the number of trees in random forests, and nA represents the number of attributes to pick randomly for each subtree in each tree in the forests. The accuracy of 97.4379% at nT=150 and nA=35

Table 1. The accuracies (%) of random forests on different nA and nT values

	nT=50	nT=100	nT=150
nA=7	96.0543	96.1503	96.2104
nA=14	96.9804	97.05	97.0972
nA=21	97.2577	97.3171	97.3345
nA=28	97.3572	97.3913	97.4167
nA=35	97.3911	97.4305	**97.4379**
nA=42	97.3718	97.3992	97.4114
nA=49	97.2840	97.3328	97.3436
nA=54	97.1686	97.2236	97.2408

was found, and this is a very good accuracy compared to the other researches. For example, J48 which is java implementation of C4.5 for Weka achieved the accuracy of 94.6366% using the same computer in 10-fold cross validation.

Table 2 shows the comparison of the result with the previous researches. As you see in the table, testing in other research used only one testing set, so that it's not as objective as 10-fold cross validation. Moreover, the results of GS SVM [13] and SVM of transformed classes [14] may not be directly comparable, because their results are based on transformed data of two classes. For more direct comparison SMO in weka was run for 90% and 10% of data for training and testing respectively. SMO is a SVM that can treat multiple classes. The accuracy is 72.6%.

Table 2. The comparison of accuracy in each data mining method

Accuracy (%)	method	training	testing
70.58	ANN	11,340 instances	565,892 instances
91.86	Scalable decision tree	60%	40%
97.4	GS SVM of class 2 only	400,000 instances	50,000 instances
95.9~99.8	SVM of transformed classes	90%	10%
94.6	J48	10-fold CV	10-fold CV
97.4	Random Forests	10-fold CV	10-fold CV

3 Conclusions

The data set of forest cover types based on cartographic data consists of very large data records. The data set covers four wilderness areas in the Roosevelt National Forest of northern Colorado, and the areas can represent forest of ecological processes because of minimal human-caused disturbances, and the data set has seven forest cover types. The size of data set is very large. It contains 581,102 instances in 54 conditional attributes, and one class attribute of forest cover types that consists of major and minor classes. Therefore, we need some clever data mining technique for the very large data. Several researches were done for better data mining results including ANN that is based on samples, scalable decision trees, and SVM that is based on class transformed data.

The proposed method based on larger nA values and smaller number of trees in our random forests than conventional random forests could find very accurate results and it is the best result according to literature survey. Because random forests can overcome the weakness of decision trees, I believe that we may find very accurate classifier of random forests for other very large data sets that have similar property like the data set, if we apply random forests similarly.

References

1. Calaway, R., Edlefsen, L., Gong, L.: Big Data Decision Trees with R, Revolution Analytics White Paper (2012), http://www.revolutionanalytics.com/why-revolution-r/whitepapers/RevoScaleRDecisionTrees.pdf

2. Quinlan, J.: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc. (1993)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth International Group, Inc. (1984)
4. Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 Algorithms in Data Mining. Knowledge and Information Systems 14, 1–37 (2008)
5. Hickey, R.J.: Structure and Majority Classes in Decision Tree Learning. Journal of Machine Learning Research 8, 1747–1768 (2007)
6. Ludermir, T.B., Yamazaki, A., Zanchettin, C.: An Optimization Methodology for Neural Network Weights and Architectures. IEEE Transactions on Neural Networks 17(6), 1452–1459 (2006)
7. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
8. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
9. Statnikov, A., Wang, L., Aliferis, C.F.: A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-based Cancer Classification. BMC Bioinformatics 9 (2008)
10. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
11. Blackard, J.A., Dean, D.J.: Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables. Computer and Electronics in Agriculture 24, 131–151 (1999)
12. Li, X.: A Scalable Decision Tree System and its Application in Pattern Recognition and Intrusion Detection. Decision Support Systems 41, 112–130 (2005)
13. Bo, L., Wang, L., Jiao, L.: Training Hard-margin Support Vector Machines Using Greedy Stagewise Algorithm. IEEE Transactions on Neural Networks 19(8), 1446–1455 (2008)
14. Trebar, T., Steele, N.: Application of Distributed SVM Architectures in Classifying Forest Cover Types. Computers and Electronics in Agriculture 63, 119–130 (2008)
15. Moore, D., McCabe, G., Duckworth, W.M., Alwan, L.: The Practice of Business Statics: Using Data for Decisions, 2nd edn. W.H. Freeman (2008)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)

Application for Temporal Analysis of Scientific Technology Information

Myunggwon Hwang[1], Do-Heon Jeong[1], Jinhyung Kim[1], Jangwon Gim[1], Sa-kwang Song[1], Sajjad Mazhar[1,2], Hanmin Jung[1], Shuo Xu[3], and Lijun Zhu[3]

[1] Korea Institute of Science and Technology Information (KISTI), 245 Daehak-ro, Yuseong-gu, Daejeon (305-806), South Korea
[2] Korean University of Science and Technology (UST), 217 Gajung-ro, Yuseong-gu, Daejeon (305-350), South Korea
[3] Information Technology Supporting Center, Institute of Scientific and Technical Information of China, No. 188 Longwan St. South, Huludao, Liaoning 125105, P.R. China
{mgh,heon,jinhyung,jangwon,esmallj,ms,jhm}@kisti.re.kr, {xush,zhulj}@istic.ac.cn

Abstract. In recent, business intelligence becomes one of important issues due to various analyses on technology trends. Especially, understanding the relations and influences between technologies is core property for the high-performed analysis. To do this, a few works have utilized ontologies constructed automatically but still have many errors and it causes difficulty while interpreting technology trends. Therefore this paper introduces an application which visualizes relationships and influences between technologies according to time series. Our application provides clues for intuitive observations of relationship change between technologies.

Keywords: Technology information, temporal analysis, business intelligence, co-occurrence.

1 Introduction

An importance of technology trend analytics is getting bigger in point of business intelligence view. The trend analytics starts from understanding technology development according to time series, and contains measuring current status and development speed, and predicting maturity level [1]. Technologies give and take some influences each other, and there is a case of that creation or development of a technology makes related ones strong or weak even disappear sometimes. For example, the technology 'speech recognition' of Nuance Communications[1] gave positive and strong influences to the 'smart phone' of Apple Inc.[2]. To understand the influences, there is a work which defines relationships (i.e. hierarchy and relations) between technologies

[1] Nuance Communications in Wikipedia: http://en.wikipedia.org/wiki/Nuance_Communications
[2] Apple Inc. in Wikipedia: http://en.wikipedia.org/wiki/Apple_Inc

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_30,

collected automatically from diverse resources, and which measures influences between the technologies [2]. However automatic ontology construction still contains many errors in the relationships [3]. Therefore, our research is focusing on the probabilistic analysis of relationships and influences according to the time series. And this paper, as a start point of the work, introduces a simple application which deals with temporal co-occurrence between technologies.

This paper is organized as follows. Section 2 describes the temporal co-occurrence of technology information. In section 3, we explain our application. Finally, we conclude this paper in the fourth section.

2 Temporal Co-occurrence of Scientific Technology Information

This section describes information of technologies and a method for temporal co-occurrence based on the data.

This work is small part of InSciTe project[3,4] and the project aims at providing technology intelligence service by utilizing patents, papers and Web documents collected from 2001 to 2011 [4]. For the temporal co-occurrence, we use only patent data and table 1 shows its statistics.

Table 1. Statistics of patent documents and technologies extracted from the documents published in each year

Year	Count of patents	Count of technologies
2001	237,590	28,389
2002	255,584	30,031
2003	269,945	31,337
2004	301,028	33,510
2005	336,937	35,844
2006	362,739	38,143
2007	390,492	40,592
2008	472,819	45,554
2009	514,077	48,057
2010	467,727	44,548
2011	343,998	37,018

For the temporal co-occurrence, we selected technologies which appeared in the same year and used Dice-Coefficience (1) for related degree between the technologies.

$$Rel = \frac{2\times|t_{i,year}\cap t_{j,year}|}{|t_{i,year}|+|t_{j,year}|} \quad (1)$$

[3] InSciTe (Intelligence in Science & Technology): `http://semantics.kisti.re.kr/`

[4] InSciTe Adaptive in Google play: `https://play.google.com/store/apps/details?id=net.xenix.inscite&feature=search_result&hl=en`

where, *Rel* represents related degree and $t_{i,year}$ means *i*-th technology appeared in the *year*. By using the patent data and the degree, we have constructed more than fourteen million pairs of co-occurrence technologies.

3 Application and Suggestions

In the previous process, we constructed fundamental data which consists of temporal co-occurrence technology pairs with related degree calculated by Dice-Coefficient. In this section, we introduce an application which visualizes year based co-occurrence technologies by order of high related degree. Figure 1 shows the application.

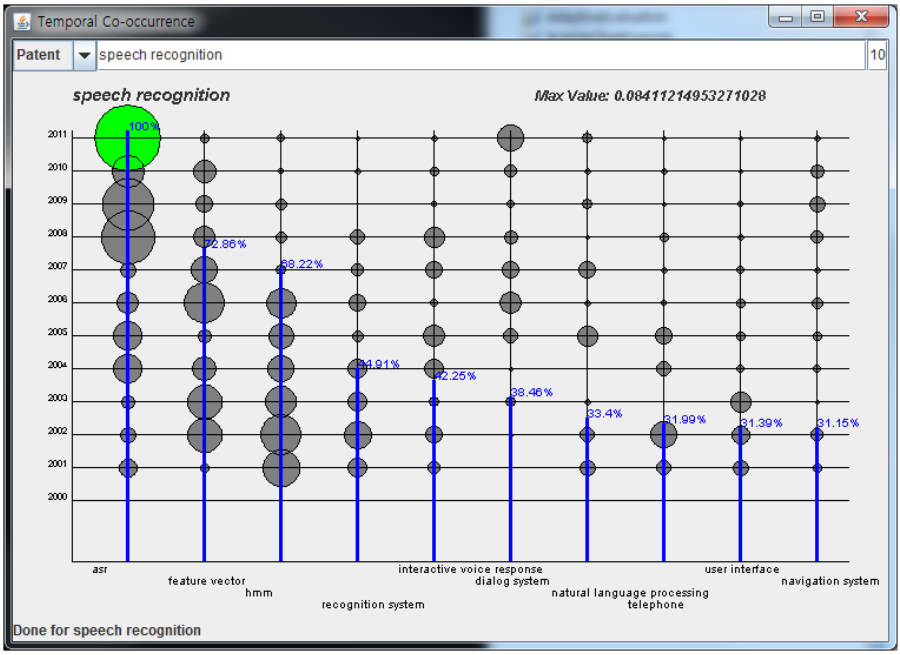


Fig. 1. Application of temporal co-occurrence

The application mainly consists of two parts: a user input panel and a dashboard for result visualization. The panel has three components again.

1. Combo box: user can select resource types (i.e. patent, paper, and Web document) but it is fixed to "Patent" until now.
2. Text field 1: users type technology in this field.
3. Text field 2: users set the maximum length of result which contains co-occurred technologies of the inserted keyword.

And the dashboard consists of x-axis, y-axis, bubble, and bar detailed as follows:

1. X-axis: technologies are positioned by order of high total sum of related degrees of all years.

2. Y-axis: bubbles are positioned on y-axis according to year.
3. Bubble: size is determined by relative related degrees in the result. And the bubble filled with green color means the maximum degree. The absolute maximum value is shown on the top right.
4. Bar: bar is colored with blue and each technology has its own bar. The bar shows relative ratio of total sum of related degrees of all years.

Figure 1 and table 2 show a result which is a set of 'patent,' 'speech recognition' and '10' in combo box, text field 1 and text field 2, respectively.

Table 2. Co-occurrence technologies of 'speech recognition'

T	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	S	R
1	0.02	0.02	0.02	0.04	0.04	0.03	0.02	0.07	0.07	0.04	0.08	0.44	1.00
2	0.01	0.04	0.04	0.03	0.02	0.05	0.03	0.03	0.02	0.03	0.01	0.32	0.73
3	0.05	0.05	0.04	0.03	0.03	0.04	0.01	0.01	0.01	0.01	0.01	0.30	0.68
4	0.03	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.01	0.01	0.20	0.45
5	0.02	0.02	0.01	0.02	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.19	0.42
6	0.00	0.01	0.01	0.01	0.02	0.03	0.02	0.02	0.01	0.02	0.03	0.17	0.38
7	0.02	0.02	0.01	0.00	0.03	0.01	0.02	0.01	0.01	0.01	0.01	0.15	0.33
8	0.02	0.03	0.00	0.02	0.02	0.01	0.01	0.01	0.00	0.01	0.01	0.14	0.32
9	0.02	0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.14	0.31
10	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.01	0.14	0.31

T: technology, '01~'11: from 2001 to 2011, S: total sum of year based related degrees, R: relative ratio, 1: asr, 2: feature vector, 3: hmm, 4: recognition system, 5: interactive voice response, 6: dialog system, 7: natural language processing, 8: telephone, 9: user interface, 10: navigation system

In the result of figure 1 and table 2, we can expect some relationship between technologies as follows.

- 'Feature vector,' 'hmm' and 'natural language processing' are considered as important element for user technology 'speech recognition.' And the current reduction of their bubble size means that they become generalized already for the user technology.
- The case of 'navigation system' ranked 10th and shows constant relationship even though its subject domain is different. 'Speech recognition' has totally 3,320 co-occurrence technologies actually. Considering both characteristics, we can expect that there is very strong connection between 'speech recognition' and 'navigtion system.'

As already shown above, the application can provide important clues to observe relationship changes and to understand current trend according to time series. The application can be considered as important factor for business intelligence. However, we cannot figure out specific relations and future trends by using only temporal co-occurrence. For the more accurate prediction, a few works should be combined such as sense disambiguation (ex. 'asr' in the result means 'automatic speech recognition') and multi-dimentional analysis methods. These are remained as future works.

4 Conclusions

In this paper we introduced an application to visualize relationships and influences between technologies according to time series. For the application, we employed patent data used for InSciTe project and measured temporal related degree based on Dice-Coefficient. As a result, we could figure out the trends of relationships and influences of technologies intuitively. We expect it can provides more effective and meaningful predictions when the application is utilized with other various analysis and visualization methods simultaneously.

References

1. Kim, J., Hwang, M., Jeong, D.-H., Jung, H.: Technology Trends Analysis and Forecasting Application based on Decision Tree and Statistical Feature Analysis. Expert Systems with Applications 39, 12618–12625 (2012)
2. Hwang, M.-N., Seo, D., Lee, S., Cho, M., Song, S.-K., Lee, J., Hong, S.-C., Choi, S.-P., Jung, H.: Ontology Model of Technical Knowledge for Analytics. In: Proceedings of International Conference on Smart Media and Applications, pp. 13–14 (2012)
3. Hwang, M.-N., Lee, S., Cho, M., Kim, S.Y., Choi, S.-P., Jung, H.: Ontology Construction of Technological Knowledge for R&D Trend Analysis. The Journal of the Korea Contents Association 12(12), 35–45 (2012)
4. Kim, J., Jeong, D.-H., Lee, D., Jung, H.: User-centered Innovative Technology Analysis and Prediction Application in Mobile Environment. Multimedia Tools and Applications (to appear)

ROI Extraction in Dermatosis Images Using a Method of Chan-Vese Segmentation Based on Saliency Detection

Zehan Wang[1], Lijun Zhu[1], and Jiandong Qi[2]

[1] Institute of Scientific and Technical Information of China (ISTIC), Beijing, China
[2] Information College, Beijing Forestry University, Beijing, China
qijiandong@gmail.com

Abstract. Extraction of ROI (Region-Of-Interest) in dermatosis images can be used in content-based image retrieval (CBIR). Image segmentation takes an important part in it. And the performance of the segmentation algorithm directly influences the efficiency of the ROI extraction results. In this paper, a method of Chan-Vese segmentation based on saliency detection to extract the ROI of the dermatosis images is proposed. Firstly the spectral residual approach (SR) [11] is used to get the saliency map of the dermatosis images. Secondly threshold segmentation is used to get the initial ROI images. Finally the Chan-Vese model is used to segment the initial ROI images to get the final ROI images, which can ensure the active contours evolve close to the object and remove the redundant information from the complex background. The experiment results show that the proposed method has the better performance than only using Chan-Vese method.

Keywords: saliency detection, Chan-Vese model, dermatosis images, ROI extraction.

1 Introduction

As one of the important human diseases, in recent years dermatosis more and more effects on human health. Skin color digital images record the information of color, texture and shape of the skin tissues which reflect the distribution of hemoglobin and subcutaneous melanin. They provide vital evidence for medical diagnosis of skin diseases. There have been methods in segmentation of skin cancer dermatoscopic images [1-2] and Psoriasis Skin Images [3-4].

Chinese Disease Question Answering System (Hestia QA) is being developed by Institute of Scientific and Technical Information of China (ISTIC). As a decision support system for dermatosis diagnosis the interface of Hestia QA is flexible and friendly which provide multiple query modes for users to choose. In image query mode people can input their images to know what kind of dermatosis they got. Certainly using the single diagnosis mode we can hardly get the accuracy diagnosis Hestia QA system combine multiple query mode to give a more accurate diagnosis.

ROI extraction techniques can detect the regions of an image which attract the attention of users. These regions usually contain the higher entropy and can represent

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_31,

for the whole image. The ROI of the dermatosis images are those diseased tissues which are distinctively different from the other normal tissues in color and texture. Many image segmentation methods are extensively used to extract the ROI of an image. As the most classic and popular region-based active contour model, Chan-Vese model [5].has been successfully used in many types of images especially the two-region images. This model is based on the Mumford-Shah functional [6] for segmentation, and is used widely in the medical imaging field, especially for the segmentation of the brain, heart and trachea [7]. However in the process of the evolution of active contour model the selection of initial contour tend to iterate from the edge of the whole image. Thus it can make the evolution result be influenced by the information of background. At the same time the time it takes is much longer.

In this article we proposed a method of ROI extraction in the dermatosis images using Chan-Vese Segmentation based on saliency detection. These images are from various dermatosis and taken by different mobile device cameras. They usually contain different degree of random noise and have difference in illumination, size, format and so on. The ROI extraction images can be used to Hestia QA image retrieval system for supporting decision.

2 The Chan-Vese Model

The Chan-Vese method [5] is inspired by the Mumford-Shan model Mumford and Shah [6] approximate the image f by a piecewise-smooth function u as the solution of the minimization problem.

Let's denote the region inside C as ω, and the region outside C as $\overline{\Omega}\setminus\omega$. Moreover, c_1 will denote the average pixels intensity inside C, and c_2 will denote the average intensity outside C.

The object of Chan-Vese algorithm is to minimize the energy function $F(c_1, c_2, C)$ .Defined by: [6].

$$F(c_1, c_2, C) = \mu \cdot Length(C) + \lambda_1 \int_{inside(C)} |u_0(x,y) - c_1|^2 dxdy + \lambda_2 \int_{outside(C)} |u_0(x,y) - c_2|^2 dxdy \tag{1}$$

Where c_1 and c_2 are constant unknowns representing the average value of u_0 inside and outside the curve, respectively. The parameters μ and λ_1, λ_2 are weights for the regularizing term and the fitting term, respectively.

Minimizing the fitting error in equation (1), the model looks for the best partition of u_0 taking only two values, namely c_1 and c_2, and with one edge C, the boundary between these two regions, given by $[u_0 \approx c_1]$ and $[u_0 \approx c_2]$. The object to be

detected will be given by one of the regions, and the curve C will be the boundary of the object. The additional length term is a regularizing term and has a scaling role.

For the curve evolution, the level set method has been used extensively, in particular where the motion is governed by mean curvature, as in [8].This formulation behaves well even with cusps, corners, and automatic topological changes. The motion by mean curvature [8] is given by

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = |\nabla \phi| div\left(\dfrac{\nabla \phi}{|\nabla \phi|}\right), \\ \phi(0,x,y) = \phi_0(x,y), t \in [0,+\infty), (x,y) \in IR^2, \end{cases} \tag{2}$$

Using the Heaviside function H and the Dirac delta function $\delta(z) = \dfrac{d}{dz} H(z)$, we can rewrite the energy function as follows:

$$\begin{aligned} F(c_1, c_2, \phi) &= \mu \int_{\Omega} \delta(\phi(x,y)) |\nabla \phi(x,y)| dxdy + \lambda_1 \int_{\Omega} |u_0(x,y) - c_1|^2 H(\phi(x,y)) dxdy \\ &+ \lambda_2 \int_{\Omega} |u_0(x,y) - c_2|^2 (1 - H(\phi(x,y))) dxdy \end{aligned} \tag{3}$$

Minimizing $F(c_1, c_2, \phi)$ with respect to the constants c_1 and c_2, for a fixed ϕ,yields the following expressions for c_1 and c_2, function of ϕ.

$$\begin{cases} c_1 = average(u_0) \quad on \quad \phi \geq 0, \\ c_2 = average(u_0) \quad on \quad \phi < 0. \end{cases}$$

Minimizing the energy $F(c_1, c_2, \phi)$ with respect to ϕ or fixed c_1 and c_2, using a gradient descent method, yields the associated Euler–Lagrange equation for ϕ governed by the mean curvature and the error terms (see [5] for more details).

3 Proposed Method

3.1 Saliency Map

For an image humans can routinely and effortlessly judge the importance of image regions, and focus attention on the important parts of them, which representative to their querying intention. And most of the remaining regions cannot be interested by them. The salient regions of an image are areas which can most attract the users' attention and represent for the content of the image. In fact because users also have the different tasks and the prior knowledge they choose the salient regions usually very subjective, getting the different regions as the salient regions of the same image.

Computationally detecting such salient image regions remains a significant goal, as it allows preferential allocation of computational resources in subsequent image analysis and synthesis. Extracted saliency maps are widely used in many computer vision applications including object -of-interest image segmentation, object recognition, adaptive compression of images, content aware image editing, and image retrieval.

The topical method based on visual feature is the saliency map method presented by Itti[9] and others. In the later study many researchers respectively presented many different saliency analysis methods. We used these methods [10-14] to get the saliency maps of our experiment images, and found that the spectral residual approach (SR) method [11] is better than the others because it can generate less noise and extract more ROI.

3.2 Chan-Vese Segmentation Based on Saliency Detection

We chose four different type dermatosis images from Google website, and then used the Chan-Vese model to segment the diseased regions and cannot get a satisfactory result. Considering this we used a method of Chan-Vese Segmentation based on saliency detection to get ROI of the dermatosis images. In the following this model will be introduced.

Step 1: Saliency detection. We used the spectral residual approach (SR) method to get the saliency maps of the images. The results are presented in Section 4.Fig.1.

Step 2: Threshold segmentation. We adopted the threshold segmentation method of the paper [11].

Given $S(x)$ of an image, the object map $O(x)$ is obtained:

$$O(x) = \begin{cases} 1 & if \quad S(x) > threshold \\ 0 & if \quad otherwise, \end{cases}$$

Empirically article [11] set $threshold = E(S(x)) \times 3$, where the $E(S(x))$ is the average intensity of the saliency map.

In fact the selection of threshold is a trade-off problem between false alarm and neglect of objects. When the threshold is smaller the whole diseased tissues can be extracted. But it also can make a lot of noises, that is, the false alarm is higher. When increasing the threshold the noise can be reduced, but it increase the neglect of objects. The trade-off problem between false alarm and neglect of objects.is that the problem between precision ratio and recall ratio.\

Step 3: Binary image filtering .In order to solve the trade-off problem we dilated the object pixels of the object map obtained by higher threshold segmentation. But while removing the noises which generated in the threshold segmentation, it also brings in some normal skin tissues which surround the diseased tissues. It makes the whole ROI extracted precisely but cannot achieve higher accuracy. Essentially speaking it cannot remove the noise. The results of the initial ROI extraction results are presented in Section 4.Fig.1.

Step 4: Chan-Vese segmentation. In order to get a better result we used the Chan-Vese model to find the accuracy contour of the diseased tissues. Considering

the accuracy of the Chan-Vese segmentation we let the zero pixel regions a pixel value $G(x)$ and got the compensated grayscale images of the initial ROI extraction results.

Given $H(x)$ of the nonzero pixel regions, the intensity of the zero pixel regions $G(x)$ is obtained

$$G(x) = N \times E\left(H\left(x\right)\right)$$

Where the $E\left(H\left(x\right)\right)$ is the average intensity of $H(x)$. The compensated grayscale images are presented in Section 4.Fig.1.

In addition, in Step 3, when dilating the binary images we adopted several approaches, and then used the disk dilation approach. In the end with different parameter combinations we got the better results of the initial ROI extraction results of the dermatosis images. These results are presented in Section 4 .Fig.1.

In Section 4.Fig.2 we presented the comparison results of the final ROI extraction, and compared the iterations and algorithm time of the Chan-Vese method and the proposed method. Apparently the proposed method can reduce the algorithm time and iterations. At the same time it can well improve the precision of the ROI extraction results.

4 Experiment Results

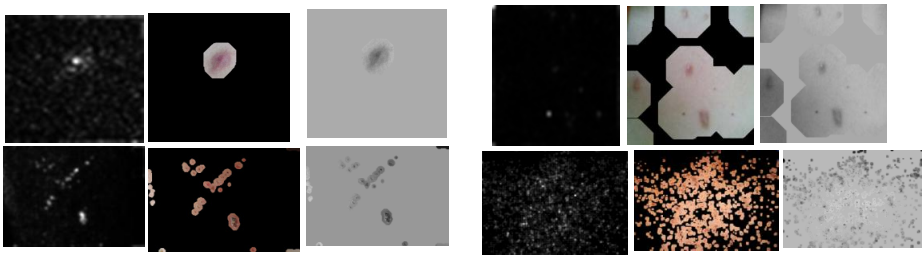


Fig. 1. (a) the saliency maps, (b) the initial ROI extraction results (c) the compensated grayscale images

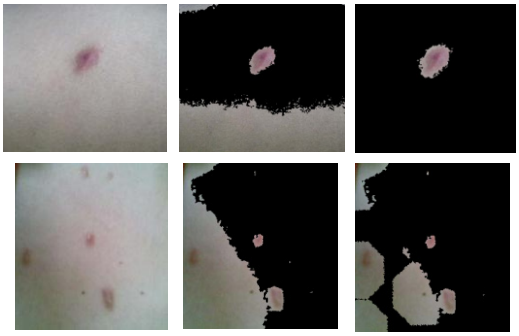
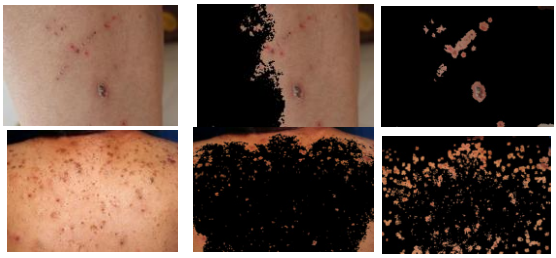


Fig. 2. (a) the original images,(b) the ROI extraction results based on the Chan-Vese method, (c)the ROI extraction results based on the proposed method



(a) (b) (c)

Fig. 2. (*continued*)

Table 1. The comparison chart of iterations

	Image 1	Image 2	Image 3	Image 4
The Chan-Vese method	160	120	800	800
The proposed method	15	85	250	150

Table 2. The comparison chart of algorithm time

	Image 1	Image 2	Image 3	Image 4
The Chan-Vese method	61.4239s	21.5153s	168.0754s	679.7835s
The proposed method	5.3006s	14.5021s	47.103s	147.083s

5 Conclusions

We discussed the application of Chan-Vese model and the proposed method in ROI extraction of the dermatosis images. With four typical dermatosis images from Google site we used several saliency detection methods to get the saliency maps, found the SR method had the advantage and then used the Chan-Vese model to get accurate results. We separately compare the iterations and algorithm time of the Chan-Vese method and the proposed method. It shows that the proposed method not only get the better ROI extraction results but can well reduce the iterations and algorithm time.

In the future work we will use more derrmatosis images and try to find an optimal parameter combination to get the more accuracy ROI extraction results. Considering the noise of the images we will use some denoising method to solve it. Even we can change the Chan-Vese model to get a more accurate result.

Acknowledgement. This research is granted by National Twelfth “Five-Year Plan” for Science and Technology Support Program: 2011BAH10B04; 2011BAH10B02.The first author would like to express gratitude and appreciation to Dr.Si Li, Dr.Ying Li and Weifeng Li for their valuable discussions.

References

1. Rahman, M.M., Desai, B.C., Bhattacharya, P.: Image Retrieval-Based Decision Support System for Dermatoscopic Images. In: 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS), pp. 285–290 (2006)
2. Mahmoud, M.K.A., Al-Jumaily, A.: Segmentation of Skin Cancer Images Based on Gradient Vector Flow (GVF) Snake. In: International Conference on Mechatronics and Automation (ICMA), pp. 216–220 (2011)
3. Taur, J.S., Lee, G.H., Tao, C.W., Chen, C., Yang, C.W.: Segmentation of Psoriasis Vulgaris Images Using Multiresolution-Based Orthogonal Subspace Techniques. IEEE Trans onSystems, Man, and Cybernetics, Part B: Cybernetics 36(2), 390–402 (2006)
4. Lu, J., Kazmierczak, E., Manton, J.H., Sinclair, R.: Automatic Segmentation of Scaling in 2-D Psoriasis Skin Images. IEEE Trans on Medical Imaging 32(4), 719–730 (2013)
5. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Processing 10(2), 266–277 (2001)
6. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. 42(5), 577–685 (1989)
7. Rousseau, O., Bourgault, Y.: Heart segmentation with an iterative Chan-Vese algorithm. University of Ottawa, Ontario (2009)
8. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulation. Journal of Computational Physics 79, 12–49 (1988)
9. Itti, L., Koch, C.: A saliency–based search mechanism for overt and covert shifts of visual attention. Vision Research 40, 1489–1506 (2000)
10. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: ACM Multimedia 2006, pp. 815–824 (2006)
11. Hou, X., Zhang, L.: Saliency detection: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
12. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1597–1604 (2009)
13. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global Contrast based Salient Region Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 409–416 (2011)
14. Hou, X., Harel, J., Koch, C.: Image Signature: Highlighting Sparse Salient Regions. IEEE Trans. on Pattern Analysis and Machine Intelligence 34(1), 194–201 (2012)

The Study on Semantic Self-sufficiency in Factual Knowledge Extraction*

Yunliang Zhang

Institute of Scientific and Technical Information of China,
15 Fuxing Road, Beijing 100038, China
zhangyl@istic.ac.cn

Abstract. In this paper, semantic self-sufficiency of elements in factual knowledge extraction is introduced to meet the real knowledge service demands. The characters of semantic self-sufficiency are analyzed, and based on which the processing strategies are proposed. Though there are still some difficulties to cope with, semantic self-sufficiency is a tool to clearly define the elements of factual knowledge and can improve the results of automatic factual knowledge extraction.

1 Introduction

In a perspective of the processing way of computer, knowledge can be divided into three categories: conceptual knowledge, linguistic knowledge and factual knowledge. The conceptual knowledge provide computer basic information of concepts, that is, what a concept is and the inner relationships between concepts. The linguistic knowledge provide computer language syntax and related things about expression and understanding, maybe including a series of rules and methods. The factual knowledge is the accumulation of the people about all kinds of everything of the world. The factual knowledge extraction and the development of factual knowledge base is the key issue to enhance the performance of question answering systems and information retrieval systems, and also important to provide high-level knowledge service.

The knowledge services need knowledge content of different granularity, such as knowledge objects and knowledge elements in patent knowledge extraction [1]. A piece of knowledge content relatively complete and independent can be called knowledge object. The major components of knowledge objects can be called knowledge elements. For example, *The electrical high volt plug connector , has a sealing unit provided between outer and inner housings that are connected with each other* is a knowledge object, *electrical high volt plug connector*, *sealing unit*

* This work is partially supported by the National Natural Science Foundation of China (Grant No. 71203208), National Science and Technology Support Program (Grant No. 2011BAH10B01) and ISTIC Key Project Program (Grant No. ZD2012-3-2).The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_32,

and *connected with each other* and so on are knowledge elements. It is very important to clearly define the elements of knowledge objects, so that the literatures would be used in a deeper level easily. The representation of knowledge objects and knowledge elements can use natural language, but not all the natural language description in a literature must be included in knowledge objects and knowledge elements. Therefore, a specific literature is written in natural language, only part of which is about the fact he want to express, the other part is about the language itself, just to make the reader understand the fact. So generally the knowledge extraction process is the separation process of factual knowledge bricks from linguistic cement. In this process, the boundaries of different knowledge objects and knowledge elements should be clarified, and all the necessary elements of the knowledge objects should be extracted without omission. So in this paper, the semantic self-sufficiency is introduced from linguistics to guarantee proper knowledge extraction. The original meaning of semantic self-sufficiency is a property that a language unit can completely and enough represent the meanings that you want to express and can be sentence elements [2, 3]. Here, it is borrowed to describe the property that a knowledge unit(knowledge element) can completely and enough represent the knowledge objects that you want to express.

Because the immaturity of semantic analysis and there are a lot of difficulties in the knowledge expression for knowledge objects. The common extraction work focus on knowledge elements that without relationships with knowledge objects. Name entities such as person, organization, place, and figures of time, amount are basic extracted knowledge elements. In ontology learning area, there are extraction work about relationships of synonyms and hierarchy. With the development of knowledge representation method such as framework and case grammar and the appearance of semantics annotation. The study about the knowledge objects and extraction of knowledge elements that under the constraints of knowledge objects is emerging. ISTIC(Institute of Scientific and Technical Information of China) starts work about standard specifications, manual tagging and automatic extraction about deep content on patent literatures. In the process, it is found that in deep content processing, the analysis and study of semantic self-sufficiency is a big problem hard to avoid. We will discuss the characteristics, processing strategies and some problems of semantic self-sufficiency in the following sections.

2 The Characteristics Analysis of Semantic Self-sufficiency

2.1 Decomposability

Knowledge elements semantic self-sufficiency in fact can be divided into two interrelated issues: that is, knowledge elements boundary judgement and knowledge elements sufficiency judgement.

Knowledge elements have no definite form, and they may be words, phrases, even sentences, knowledge elements boundary judgement then become a task in three level:word boundary judgement, phrase boundary judgement, sentence boundary check. The boundary words, phrases and sentences is form boundary, but the boundary of

knowledge elements is semantic boundary. There are some relationship between semantic boundary and form boundary, that is a semantic boundary must be a form boundary, but not every form boundary is a semantic boundary. So all form boundaries are candidate semantic boundaries. For example, *high volt* in section 1 is a phrase but not a knowledge element, in fact it is a part but not a necessary part of a knowledge element. The boundaries of *high volt* are form boundaries rather than semantic boundaries. Relatively, The form boundary analysis technologies is mature in natural language processing, so the following methods can be adopted: the check of form boundaries of knowledge elements is a prior task of the check of semantic boundaries and decide the border in the knowledge elements sufficiency check step.

Semantic sufficiency check is to judge if the boundaries of knowledge elements in a larger view(knowledge object)is reasonable or not. The check needs to construct a series of standards about a specific knowledge object contains what kinds of knowledge elements and the amount and constraints of every kind of knowledge element.

2.2 Elements Finiteness

Knowledge base of knowledge from the various types of literatures, and in a specific literature, such as a patent, the description of knowledge objects is not that complicated. Usually in a non-nested knowledge object, the amount of knowledge elements involved will not be more than 5, it is also accord with the 7 ± 2 theory in psychologies. In fact, in the most common circumstances, there are 3 knowledge elements in a knowledge object, which is very similar with triples in resources description framework(RDF). To know the finiteness of knowledge elements in a knowledge object is very important for the preliminary sufficiency check. If there are more than 5 knowledge elements in a knowledge object, it need a careful check. There maybe 3 cases: 1) There is element nested with other knowledge object; 2)There is mistake about boundaries of some knowledge elements; 3)It is correct in some special expression, but it is rare to appear.

2.3 Elements Elasticity

To keep the semantic self-sufficiency of knowledge elements, we should separate the limited amount of knowledge elements with correct boundaries, but it is not a single solution problem. For that different authors have different writing habits. Some authors use concise language, but others usually use a lot of qualifiers. So the elasticity of knowledge elements about semantic self-efficiency is that each value of boundaries in a discrete space can satisfy the self-sufficiency, and the value can be a value between the lower limit and upper limit. The elastic lower limit is to keep the core content of the semantic framework, and the elastic upper limit is to keep all the qualifiers of the authors, and there are some intermediate situations with part of all qualifier of authors. For example, *The electrical high volt plug connector* in section 1 is the upper limit and *The connector* is the lower limit, *The plug connector*, *The high volt plug connector* are also legal knowledge elements. In extraction practice, it is more easy to adopt either the lower limit or the upper limit.

3 The Processing Strategies of Semantic Self-sufficiency

3.1 Element Boundary Check

In general, there are two categories of methods to check the boundaries of elements, that is, through internal coupling relations approach and language-specific symbols approach. Internal coupling relationship approach is from the view that if the adjacent parts should be combined to a element, and language-specific symbols approach is from the view that if language symbols are separators of different knowledge elements. The two approaches are from different philosophical thought of combination and separation, and can be used together.

The check method varies with the demand about the elasticity. For the elastic lower limit, the key problem is to find the core of a knowledge element. For different language, the pattern are different, for Chinese and English without subordinate clauses, the core usually in the rear of a knowledge element. But this law is not effective for complicated situation, such as prepositional phrases, nested sentences and inversion sentences. For the elastic upper limit, the methods must cover the three situations about knowledge elements of words, phrases and sentences. Different forms knowledge elements have different laws. For words, dictionary match and unlisted word discovery can be combined. For phrases, we can use the composition rules of phrases and semantic relevance computation. The composition rules use the information of part of speech, the concept categories and so on to check the semantic self-sufficiency qualitatively. The semantic relevance use the quantitative value to check the relevance of neighbour words. For sentences, it is relatively easy for that there are explicit punctuation characters to separate different sentences. It must be pointed that the semantic self-sufficiency is not a separate concept, it is meaningful only in the scope of a knowledge object.

3.2 Qualitative Elements Constraints Check

The constraints of semantic self-sufficiency is divided into two kinds, that is, categories constraints and relationships constraints. The categories constraints require a knowledge element in a specific knowledge object must belong to one or more categories. For example, a knowledge object in *vehicle* domain about structural composition are limited to *parts* and *complete machine*. The relationships constraints require knowledge elements categories in a specific knowledge object must has accord some rules. For example, in a case about *replacement* in technology, the substitutions must be the same categories, for example, both are *parts* or both are *substances*, you can't replace a *part* with a *substance*.

3.3 Quantitative Elements Constraints Check

Some research in natural language processing, such as Hierarchical Network of Concepts (HNC) theory provide semantic sentence categories, with basic and hybrid sentence categories, all the natural language expressions can be classified into basic or

hybrid sentence categories [4,5]. Every sentence category can be activated by the verb (Of course, the sentence category must be verified by a series of process, but the related theories and tools are developed by other researchers.) And every sentence category has the constraints of how many knowledge elements involved. For example, *replace* sentence category has two knowledge elements except for the verb element, and the *transfer* sentence category has three knowledge elements except for the verb element. HNC theory also has the technologies about boundary analysis for semantic chunks, which can be utilized in knowledge elements boundary analysis.

4 Conclusion

Factual knowledge is a very important base for knowledge service, and the factual knowledge extraction should consider both the knowledge elements themselves and the knowledge objects and their constraints on knowledge elements. In this paper, we borrowed semantic self-sufficiency to study the constraints of knowledge elements from knowledge objects. The semantic self-sufficiency has the characteristics of decomposability, elements finiteness and elements elasticity. The main strategies for dealing with self-sufficiency are boundary check, qualitative and quantitative constraints check for knowledge elements. There are some difficult tasks of nested structure analysis, inverted word order, coreference analysis and omission recovery etc. will be studied in the future study.

References

1. Zhang, Y.L., Gui, J., Zhu, L.J., Qiao, X.D.: The Development of Deep Content Indexing Standard of Chinese Patent. Digital Library Forum 11, 18–21 (2008)
2. Lin, Y.H.: Basic Direction Word Study from Weijin to Tang, PhD thesis, Huazhong University of Science and Technology (2006)
3. Wu, J.S.: Corpus Driven Study on Chinese-English Corresponding Unit Transition Analysis, Master thesis, Henan Normal University (2007)
4. Huang, Z.Y.: The Hierarchical Network of Concepts theory. Tsinghua University Press, Beijing (1998)
5. Huang, Z.Y.: The Fundamental theorem and mathematic physics expression of the language concept space. Ocean Press, Beijing (1998)

XML-Based Document Retrieval in Chinese Diseases Question Answering System*

Haodong Zhang[1,2], Lijun Zhu[2], Shuo Xu[2], and Weifeng Li[2]

[1] Network Center, Science and Technology Daily, Beijing, China
hdzhang06@126.com
[2] Institute of Scientific and Technical Information, Beijing, China
zhulj@istic.ac.cn, pzczxs@gmail.com, zhichen08@163.com

Abstract. A Chinese Diseases Question Answering System(Hestia QA） is being developed by ISTIC. As a part of Hestia QA, a XML-based document retrieval and similarity calculation model is established here. The texts which describe diseases in Chinese are indexed and wrapped in XML tags. The query is compared with related tags in XML document and the similarity is calculated with a deformed cosine similarity algorithm. The Chinese terms semantic similarity calculation algorithm is used to get the similarity of two terms in the system. The result shows that our model works well. The Chinese disease XML datasets will be analyzed in different granularity levels or dimensions. The corpus of diseases in Chinese will be established after the automatic XML annotation software is completed in the next step.

Keywords: ML, Chinese Diseases QA, Chinese Terms Similarity, Cosine Similarity.

1 Hestia Question Answering System

A Chinese Diseases Question Answering System（Hestia QA）is being developed by Institute of Scientific and Technical Information of China(ISTIC). As a part of Hestia QA, a XML-based document retrieval and similarity calculation model is established here. Some retrieval systems are over relational databases, whose advantages are that from the point of view of the software development, the interface between users and a relational database is flexible and friendly. However, there are also disadvantages to have relational database. The cost of the data pre-processing is high. When the system requires a frequent change for the data, the cost will be high and it is also difficult to have a change. In this article, a semi-structured XML data retrieval system will be established. Semi-structured datasets with domain-specific semantic annotation can be rapidly formed with an automatic annotation system in the future.

* This research is granted by National Twelfth “Five-Year Plan” for Science and Technology Support Program: 2011BAH10B04.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_33,

2 Related Work

It is an important step from keywords matching to semantic retrieval in information retrieval. The phrase-based statistic has a positive but limited help to improve the performance of text retrieval and linguistic analysis is not better than phrase-based statistic [1]. In the following, one semantic computation model is introduced.

2.1 Related Rext Retrieval Model

Semantic Tree Model

Usually, the query and text are decomposed into words and words are represented as vectors. It is a common way to use the simple data structure tree to show the relationship between words, which is called the semantic tree model [2].

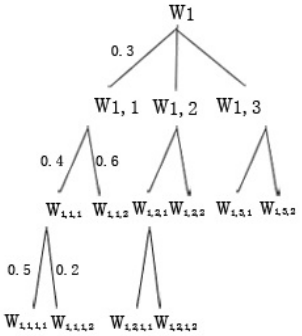


Fig. 1. Semantic Tree model

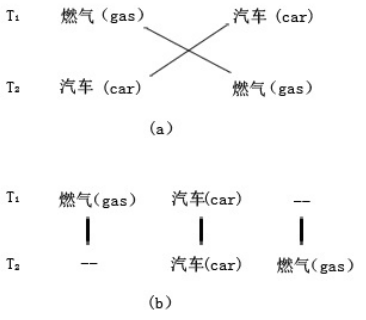


Fig. 2. The correspondence between T_1 and T_2

In a tree（Fig. 1）, between the node w_i and its child node w_j , the path weight is their similarity, e.g. $Sim(w_1,w_{1,1})$=*weigh-of-path*$(w_1,w_{1,1})$=0.3.

2.2 XML Documents in Hestia QA

It is a popular method in text retrieval to convert texts into XML documents and carry on data indexing and processing, e.g. [5]. Hestia QA describes the disease in text. In this article, the texts are indexed and marked up in XML. For example,

“毛囊炎是局限于毛囊口的化脓性炎症”

(“Folliculitis is suppurative inflammation in follicular orifice.”)

The note is wrapped in XML tags as follow,

<diseaseName>毛囊炎(folliculitis) </diseaseName>是局限于

<bodyParts>毛囊口(follicular orifice)</bodyParts >的

<diseaseName>化脓性炎症(suppurative inflammation)</diseaseName>

3 Tongyici Cilin and Chinese Terms Semantic Similarity

3.1 Tongyici Cilin

Tongyici Cilin(*Cilin 1*)[6] is a Chinese thesaurus by Mei Jiaju, et al in 1983. *Cilin* has a vocabulary of 53859 terms and all in a hierarchy tree. The vocabulary is divided into large, medium, small classes. There are 12 big classes (capital letter), 94 medium classes (small letter) and 1428 small classes (two-digit decimal integer). Afterwards, Information Retrieval Lab of Harbin Institute of Technology compiled *Tongyici Cilin* extended version (*Tongyici Cilin 2*) [7]. There are 77343 terms. *Tongyici Cilin 2* has a structure of five hierarchies, including big class, medium class, small class, synset and subsynset (a line). The forth level is called synset, corresponding to every paragraph of the third level in *Cilin 1* and the fifth level is subsynset, corresponding to every line of the third level in *Cilin 1*. The eighth bit is a mark, including =, # and @. = stands for "equal", "synonymous". # represents "similar", "related". @ is "self-enclosed", "independent" and doesn't have synonym or related words in the dictionary. For instance, Aa01A04= is {劳力 劳动力 工作者}(The three Chinese phrases are similar to the word labour).

Table 1. Word code in *Cilin 2*

code bit	1	2	3	4	5	6	7	8
Symbolic	A	d	1	2	B	0	2	=,#,@
Classification	big class	medium class	small class		Synset	Subsynset		
Hierarchy	1	2	3		4	5		

3.2 Chinese Terms Semantic Similarity Calculation

There are two terms, T_1=<$PT_{1,1}$,$PT_{1,2}$,…, $PT_{1,m}$>, T_2=<$PT_{2,1}$,$PT_{2,2}$,…, $PT_{2,n}$>. The similarity of T_1 and T_2 is calculated [8].

Example 1[8] T_1=<燃气(gas), 汽车(car)>, T_2=<汽车(car), 燃气(gas)>. According to the correspondence in Fig. 2 (a), the similarity of T1 and T2 is 1.0, which is unreasonable, because the sequence of Chinese terms is not considered. In Chinese, the order is important to semantics.

$$Sim(T_1,T_2) = 0.3\times\left(\frac{1}{2}+\frac{1}{2}\right)\times 2 + 0.2 + \frac{2}{2}\times\left[\left(\frac{1}{1+2}+\frac{2}{1+2}\right)\times 1+\left(\frac{2}{1+2}+\frac{1}{1+2}\right)\times 1\right] = 1.0$$

The corresponding relationship in Fig. 2 (b) is generated by the algorithm [8]. Following the correspondence, the result is 0.5, which is more reasonable.

$$Sim(T_1, T_2) = 0.3 \times \left(\frac{1}{2} + \frac{1}{2}\right) \times 1 + 0.2 \times \frac{2}{2} \times \left(\frac{1}{1+2} + \frac{2}{1+2}\right) \times 1 = 0.5$$

4 Document Retrieval Based on XML

4.1 Introduction

As a part in Hestia QA, the document retrieval is based on XML. One line includes two input fields. First, there is a drop-down list, which has six options, Disease Name, Body Parts, Patient Description, Symptom, Colour and Skin Appearance. Then, the description is input in the second text field. For example, select Colour in the drop-down list and input 红(red) for description. Now there are 4 lines and user can input 1-4 lines. After the query is submitted, the input will be compared with the XML documents and calculate the similarity. For instance, the input is <Colour, 红>. The system will search all the <colour></colour> tags in the XML document and compare their value with "红". Then the similarity of the input and the XML document is calculated with the algorithm.

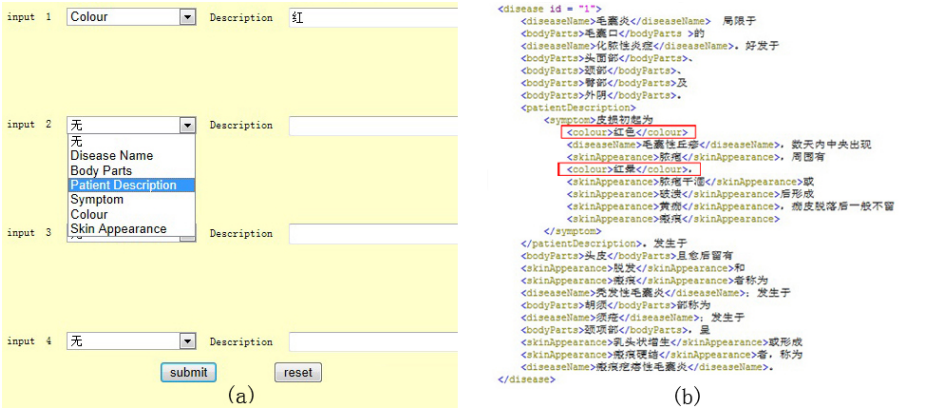


Fig. 3. The input matches XML tags

4.2 Algorithm and Implementation

Now, 1-4 items can be entered. Fig. 3. (a). $input=\{input_1, input_2, \cdots input_i\}$, $\{i | 1 \leqslant i \leqslant 4, i \in Z\}$. In $input_i$, A is the attribute from the drop-down list and D represents the description from the text input field, e.g. $input_i$=<colour,红(red)>, A=*colour*, D=红. The tags related to A are searched and their values form a set $\{V_1, V_2, \ldots, V_n\}$. According to Chinese terms semantic similarity calculation algorithm

[8], the similarity of D and V_n is $Sim(D,V_n)$. The sum of the similarities is $SSim = \sum_{j=1}^{n} Sim(D,V_j)$. The similarity of D and $\{V_1,V_2,\ldots,V_n\}$ is,

$$Sim(D,V) = Sim(D,\{V_1,V_2,\ldots,V_n\}) = \frac{SSim}{n} = \frac{\sum_{j=1}^{n} Sim(D,V_j)}{n} \quad (2)$$

Therefore, the similarity of one item $iniput_i$ and a XML document is $Sim(D,V)_i$. There may be 1-4 input items. We have vector Vec_1 =($Sim(D,V)_1$, $Sim(D,V)_2$,$\cdots$,$Sim(D,V)_i$) and Vec_2=(1,1,$\cdots$, 1). All the elements in Vec_2 are 1. The length of Vec_2 is i, which always equals to Vec_1. According to the algorithm of cosine similarity [9], the similarity of Vec_1 and Vec_2,

$$Sim(Vec_1,Vec_2) = \frac{Vec_1 \cdot Vec_2}{\| Vec_1 \| \| Vec_2 \|} \quad (3)$$

To compute the similarity of *input* and a XML document, $Sim(Vec_1,Vec_2)$ will be multiplied by the corresponding coefficients. When $input_i$ is matched with a XML document *Doc*, the terms which are similar to D in $\{V_1,V_2,\ldots,V_n\}$ is m. ($Sim(D,V_n) \neq 0$)

$$Sim(input,Doc) = Sim(\{input_1,input_2,\ldots,input_i\},Doc) = Sim(Vec_1,Vec_2) \times \prod_{j=1}^{i} \frac{SSim_j}{m_j} \quad (4)$$

When $input_i$ matches a XML document, if m=0, it means the description D is not similar to any related tags. Then, $SSim_i/m_i$=0.2.

Example 2 $input_1$=<*colour*, 红 （ red） >, $input_2$=<*Body Parts*, 头皮 (scalp)>. Calculate the similarity of the two input items and the XML document in Fig. 3(b).

$$Sim(\text{红},V)_1 = \frac{0.45+0.45}{2} = 0.45$$

$$Sim(\text{头皮},V)_2 = \frac{0.25+1}{8} \approx 0.1563$$

Then, Vec_1=(0.45,0.1563), Vec_2=(1,1)

$$Sim(Vec_1,Vec_2) = \frac{0.45 \times 1 + 0.1563 \times 1}{\sqrt{1^2+1^2} \times \sqrt{0.45^2+0.1563^2}} \approx 0.8999$$

$$Sim(input,Doc) = Sim(Vec_1,Vec_2) \times \frac{0.90}{2} \times \frac{1.25}{2} \approx 0.2531$$

5 Conclusion and Future Work

Input some queries into the system. $input_1$=<*Colour*,红(red)>, $input_2$=<*Body Parts*,头皮(scalp)> and $input_3$=<*Disease Name*, 须疮(scabies)>.

Table 2. Result

id	Disease name	similarity
1	毛囊炎(folliculitis)	0.2487
2	疖毛囊深部及周围组织的化脓性炎症(Purulent inflammation)	0.0151
3	多个相邻毛囊及毛囊周围炎症相互融合而形成的皮肤深层感染(Skin infection)	0
4	丹毒多由乙型溶血性链球菌感染(streptococcal infection)	0.0455
5	疥疮(scabies)	0.0191

{$input_1$, $input_2$, $input_3$} is more similar to the document of folliculitis. As shown in Table 2, the input has a similarity of 0.2487 with folliculitis, which is much higher than others. Therefore the model in this article is reasonable.

In this model, the texts which describe the disease are wrapped in XML tags and converted to XML documents. The input items are compared with their related tags in XML documents. The Chinese terms semantic similarity calculation based on pairwise sequence alignment is used to calculate the terms' similarity. The deformed cosine similarity is applied to the computation when compare the input items and a XML document. It is platform free and more compatible. Compared with systems based on relational databases, it reduces the servers' load.

Until now, the texts are wrapped in XML tags manually, so there has not been a corpus of diseases in Chinese. An automatic XML annotation system is being established. After enough XML documents are indexed, the statistics of different tags' weight will be studied when calculating the similarity of the input items and XML documents. For example, the weight of colour tags is α and the weight of symptom tags is β. The value α/ß will be analyzed to help calculating the similarity. Then the algorithm will be improved.

After the text documents are marked with XML tags, they formed a XML datasets with domain-specific semantic annotation. With different tags, the XML dataset can be clustered, queried and analyzed in different granularity levels or dimensions, e.g. {(Disease Name), (Disease Name, Symptom), (Disease Name, Symptom, Colour)}.The XML dataset becomes a kind of data cube [9]. *count()* , *max()* and other measures are able to apply to the XML dataset. There can be multidimensional data analysis on the XML dataset as cube space. Complex aggregation or other cube space explorations will be used to the XML dataset which describes diseases in Chinese. It is the next step.

References

1. Zhao, J., Jin, Q.-L., Xu, B.: Semantic Computation for Text Retrieval. Chinese Journal of Computers 28(12) (December 2005)
2. Jin, Q.-L., Zhao, J., Xu, B.: Query expansion based on term similarity tree model. In: Proceedings of the International Conference on Nature Language Processing and Knowledge Engineering (NLPKE), Beijing, 400-406 (2003)
3. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
4. Church, K.W., Gale, W.A.: Inverse document frequency (IDF): A measure of deviations from Poisson. In: Proceedings of the 3rd Workshop on Very Large Corpora, Boston, MA, USA, pp. 121–130 (1995)
5. Jiang, T.: Research on Rich-text XML Document Retrieval. Jiangxi University of Finance and Economics (2006)
6. Mei, J.J., Zhu, Y.M., Gao, Y.Q., Yin, H.X.: Tongyici Cilin: Shanghai Lexicographical Publishing House, Shanghai, China (1983) (in Chinese)
7. Tongyici Cilin (Extension Edition), `http://www.irlab.org`
8. Xu, S., Zhu, L., Qiao, X., Xue, C.: A Novel Approach to Chinese Terms Semantic Similarity Calculation Based on Pairwise Sequence Alignment. Journal of the China Society for Scientific and Technical Information 29(4), 701–708 (2010)
9. Han, J., Kamber, M., Pei, J.: Date Mining Concepts and Techniques (March 2012)

Mathematical Document Retrieval Model Using Structural Information of Equations in Pseudo-documents

Yeongkil Song, Junsoo Shin, and Harksoo Kim

Program of Computer and Communications Engineering, College of Information Technology, Kangwon National University, Republic of Korea
{nlpyksong,nlpsjs,nlpdrkim}@kangwon.ac.kr

Abstract. Math-aware search engines are required to effectively retrieve mathematical documents including various equations. In this paper, we propose a mathematical document retrieval system by which users can retrieve documents using any combination of keywords and equations. The proposed system indexes equations and their surrounding keywords from mathematical documents. Then, it searches and ranks mathematical documents using a language model modified for the heterogeneous indexing units (i.e., mixtures of equations and keywords). In the experiments, the proposed system performed well, especially for high ranks.

Keywords: Mathematical Document Retrieval, Heterogeneous Indexing Term, Pseudo-document.

1 Introduction

Many online documents in science, engineering, and mathematics include various equations. In this paper, we assume that equations are represented in mathematical markup language (MathML) [3]. To effectively retrieve such documents, some math-aware search engines have been developed [1,4,6]. The previous math-aware search engines were focused on methods to group similar equations by using regular expressions or extracting index terms such as function names, variables, and constants. However, these systems used handcrafted rules for indexing the structural information of equations. Moreover, they did not consider the relations between equations and their surrounding keywords. Many users want to search for structurally similar equations although the variables in the equations may be quite different from each other. In addition, they want to search documents by keywords although the retrieval targets are equations. To satisfy the needs of these users, we propose a mathematical document retrieval system to simply index two kinds of heterogeneous terms (i.e., equations and keywords) within the same framework.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_34,

2 Mathematical Document Retrieval Model

2.1 Extraction of Indexing Terms

To extract indexing terms from mathematical documents, the proposed system constructs pseudo-documents. A pseudo-document consists of three paragraphs. The first is the group of all sentences between the previous equation and the retrieval target equation, the second is the retrieval target equation, and the last is the group of all sentences between the retrieval target equation and the next equation. Then, the proposed system extracts keyword indexing terms and equation indexing terms from each pseudo-document. By using a conventional morphological analyzer, the proposed system extracts content words (i.e., nouns, verbs, adjectives, and adverbs) as keyword indexing terms. Table 1 lists the three types of symbolic information that the proposed system extracts as equation indexing terms by using a MathML parser.

Table 1. Symbolic information in equations

Name	Meaning	Example
STRUCTURE	The structure tags	msqrt, mroot
OPERATOR	The operators	+, -, *, /
VARIABLE	The variables	*x*, *y*, *z*

Table 2 lists the 16 structure tags that the proposed system extracts from MathML equations.

Table 2. Structure tags and their meanings

Tag	Meaning	Tag	Meaning	Tag	Meaning
mtable	A matrix	msub	A_{2}	mfrac	$\frac{A}{B}$
mtr	A row in a matrix	msubup	$\int_{a}^{b}$	mo	An operator
mtd	A column in a matrix	mover	$\overline{A}$	mn	A constant
mroot	$\sqrt[3]{A}$	munder	$\lim_{a \to 0}$	mi	A variable
msqrt	$\sqrt{A}$	munderover	$\sum_{a=0}^{b}$		
msup	A^{2}	mfenced	$(A+B)$		

Table 3 shows an example of symbolic information that the proposed system extracts from MathML equations. As shown in Table 3, $(a_x+b)^x$ and a_x+b have the same operator and the same variables. To discriminate these kinds of equations, the proposed system extracts structure tags as additional symbolic information.

Table 3. An example of symbolic information

Equation	STRUCTURE	OPERATOR	VARIABLE
$(a+b)^x$	mi, mo, mfenced, msup	+	*a*, *b*, *x*
$ax+b$	mi, mo	+	*a*, *b*, *x*

2.2 Ranking of Pseudo-documents

To rank pseudo-documents, the proposed system uses a language model [5], as shown in Equation (1).

$$\log p(q \mid \tilde{d}) = \alpha \left\{ \prod_{i:tf(q_i,\tilde{d}_k)>0} \lambda_1 \frac{tf(q_i,\tilde{d}_k)}{|\tilde{d}_k|} + (1-\lambda_1)\frac{tf(q_i,\tilde{d}_k)}{|C_k|} \right\} + (1-\alpha) \left\{ \prod_{i:tf(q_i,\tilde{d}_e)>0} \lambda_2 \frac{tf(q_i,\tilde{d}_e)}{|\tilde{d}_e|} + (1-\lambda_2)\frac{tf(q_i,\tilde{d}_e)}{|C_e|} \right\} \quad (1)$$

In Equation (1), $\tilde{d}_k$and $\tilde{d}_e$ are a set of keyword indexing terms and a set of equation indexing terms, respectively, in a pseudo-document $\tilde{d}$. The term frequency of q_i in $\tilde{d}$ is denoted by $tf(q_i, \tilde{d})$. Further, C_k and C_e are collections of keyword indexing terms and equation indexing terms, respectively, while λ_1 and λ_2 are the Dirichlet priors for keyword indexing terms and equation indexing terms, respectively [7]. The Drichlet priors are calculated according to

$$\lambda_1 = \frac{\mu_k}{|\tilde{d}_k| + \mu_k}, \ \lambda_2 = \frac{\mu_e}{|\tilde{d}_e| + \mu_e} \quad (2)$$

where μ_k and μ_e are Dirichlet smoothing parameters for keyword indexing terms and equation indexing terms, respectively. In Equation (1), α is the weighting parameter for combining two language models; a language model based on keyword indexing terms and a language model based on equation indexing terms. To set μ_k, μ_e and α, we evaluated the proposed system by using equations and names describing the equations (e.g., a cosine similarity and a language model) that are not included in a test query set. Based on this parameter calibration process, we set μ_k, μ_e, and α to 300, 800, and 0.6, respectively. The generation probability of equation indexing terms, $p(q_i \mid C_e)$, is calculated according to

$$p(q_i \mid C_e) = \frac{|C_s|}{|C_e|} p(q_i \mid C_s) + \frac{|C_o|}{|C_e|} p(q_i \mid C_o) + \frac{|C_v|}{|C_e|} p(q_i \mid C_v) \quad (3)$$

where C_s, C_o, and C_v are collections of three types of symbolic information (structure tags, operators, and variables) that compose equation indexing terms.

3 Evaluation

We collected 500 mathematical documents from arXMLiv [1] by randomly. Then, we automatically constructed 19,193 pseudo-documents from the collected documents. Next, by looking up the other documents in arXMLiv, we manually constructed a query set that consists of 50 pairs, each an equation and a name describing the equation, because we could not find any test collection for mathematical document retrieval. Finally, we had one graduate student and two university students judge the relevance of top 10 retrieved documents.

The first experiment was performed to evaluate the effectiveness of pseudo-documents by using the names describing equations as queries. Table 4 shows the differences in performance according to the document unit (full documents or pseudo-documents).

Table 4. Performances according to document units

Document Unit	P@1	P@3	P@5	P@10
Full documents	0.30	0.17	0.20	0.1
Pseudo-documents	0.40	0.23	0.20	0.1

As shown in Table 4, the system based on pseudo-documents performed better than that based on full documents, at least for high ranks.

The second experiment compared the performance of the proposed system with a previous system by using equations as queries. Table 5 shows the differences in performance.

Table 5. Comparison of performance with the previous system

Model	P@1	P@3	P@5	P@8	P@10
Youssef	0.65	0.50	0.42	0.37	0.34
Proposed system	0.69	0.53	0.42	0.38	0.34

In Table 5, Youssef indicates an equation retrieval system in which function names, operators, and variables are indexed without structure tags [6]. As shown in Table 5, the proposed system somewhat outperformed the Youssef system, at least for high ranks. As an additional experiment, we randomly chose five equations from the indexed pseudo-documents and changed variables in the equations, as shown in Table 6.

Table 6. Equations in which variables are changed

Original Equation	Modified equation
$G(\vec{z}x-y\vec{z},z)=\frac{1}{\sqrt{4\pi z}}e^{-\frac{\|x-y\|^2}{4z}}$	$G(\vec{z}a-b\vec{z},c)=\frac{1}{\sqrt{4\pi c}}e^{-\frac{\|a-b\|^2}{4c}}$
$(\mu_1-\mu_2^{\hat{}})^{-n(n-1)}\frac{1}{\Delta(\hat{q_1})\Delta(\hat{q_2})}e^{\frac{i}{2}Tr(\omega_1+\omega_2^{\hat{}})Tr(\hat{p_1}+}$	$(\alpha_1-\alpha_2^{\hat{}})^{-m(m-1)}\frac{1}{\Delta(\hat{p_1})\Delta(\hat{p_2})}e^{\frac{i}{2}M(\omega_1+\omega_2^{\hat{}})Tr(\hat{p_1}+\hat{p}$
$S\mapsto\begin{pmatrix}1_n & 0\\ 0 & A\end{pmatrix}S\begin{pmatrix}1_n & 0\\ 0 & B\end{pmatrix}$	$S\mapsto\begin{pmatrix}1_m & 0\\ 0 & \alpha\end{pmatrix}S\begin{pmatrix}1_m & 0\\ 0 & \beta\end{pmatrix}$
(n^2-16)	(A^2-16)
$w=\sum_g D^g T_g$	$w=\sum_g D^g T_{g'}$

Table 7. Comparison of rankings with the previous system

Equation in Table 6	Ranking in Youssef	Ranking in the proposed system
The 1st Eq.	253	43
The 2nd Eq.	4	1
The 3rd Eq.	1	1
The 4th Eq.	2094	189
The 5th Eq.	6	38
Average ranking	**471.6**	**54.4**

Then, we input the modified equations as queries and computed the rankings in which the original equations are retrieved, as shown in Table 7. These experimental results show that structure tags are important clues in equation retrieval.

4 Conclusion

We proposed a mathematical document retrieval model to index heterogeneous terms within a single framework. To associate equations with their surrounding keywords, the proposed system divides ordinary documents into pseudo-documents. Then, it indexes partial equations and their surrounding keywords from the pseudo-documents. In the experiments, we found that the proposed methods are effective in mathematical document retrieval. In the future, we will more intensively evaluate the proposed system with various test queries. In addition, we will perform a study on keyword weighting methods in which relations between equations and keywords are considered.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A4A01005074). This study was also supported by 2013 Research Grant from Kangwon National University(No. C1009876-01-01).

References

1. Addel, M., Cheung, H.S., Khiyal, S.H.: Math go! Prototype of a content based mathematical formula search engine. Journal of Theoretical and Applied Information Technology 4(10), 1002–1012 (2008)
2. arXMLiv, http://arxmliv.kwarc.info/files/math-ph/papers
3. MathML, http://www.w3.org/Math
4. Misutka, J., Galambos, L.: Extending full text search engine for mathematical content. In: DML 2008 Workshop, pp. 55–67 (2008)
5. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: ACM SIGIR, pp. 275–281 (1998)
6. Youssef, A.S.: Relevance ranking and hit description in math search. Mathematics in Computer Science 2(2), 333–353 (2008)
7. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems 22(2), 179–214 (2004)

Lexical Feature Extraction Method for Classification of Erroneous Online Customer Reviews Based on Pattern Matching

Maengsik Choi, Junsoo Shin, and Harksoo Kim

Program of Computer and Communications Engineering, College of Information Technology, Kangwon National University, Republic of Korea
{nlpmschoi,nlpsjs,nlpdrkim}@kangwon.ac.kr

Abstract. In morpheme-based languages such as Korean and Japanese, spacing and spelling errors that frequently occur in online documents make it difficult to reliably extract informative lexical clues for sentiment analysis. To overcome this problem, we propose a simple, reliable lexical feature extraction method for sentiment classification systems; this method targets online customer reviews in Korean, which include numerous spacing and spelling errors. The proposed method performs longest-matching between input sentences and two kinds of patterns (spacing-unit patterns and phoneme patterns) that are automatically constructed from a large POS tagged corpus. Thereafter, the method returns content words associated with the longest matched patterns. In the experiments on sentiment classification, the proposed method outperformed previous lexical feature extraction methods, which are based on conventional morphological analyzers.

Keywords: Lexical feature extraction, spacing and spelling errors, spacing-unit pattern, phoneme pattern.

1 Introduction

Most online customers carefully study previous customers' reviews associated with the products that they intend to buy before purchasing those products. There is no doubt that it is time-consuming to read numerous reviews. To reduce this time-consuming job for customers, many studies of automatic sentiment classification have been performed [1, 2]. Previous studies generally used machine-learning methods based on linguistic features that were extracted using linguistic analyzers such as morphological analyzers and part-of-speech (POS) taggers. However, most online customer reviews included numerous spacing errors (especially, in agglutinative languages) or spelling errors, which were intentionally generated to reduce the length of sentences or to express familiarity. These errors decreased the performance of the linguistic analyzers and resulted in low performance for the sentiment classifiers [3]. To resolve this problem, many text normalization methods for inflected languages (e.g., English, Spanish, German, and Norwegian) have been studied [4, 5]. However,

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_35,

these methods might not be proper for agglutinative languages such as Korean, Japanese, Turkish, and Hungarian, because they are based on string comparisons of spacing units (i.e., comparing erroneous word input with correct word entries in a dictionary). The spacing unit of agglutinative languages is morphologically more complex and usually longer than that of inflected languages. Therefore, the text normalization methods in agglutinative languages are generally performed based on morphemes not spacing units because string comparison based on spacing units require large memory. However, conventional morphological analyzers in erroneous sentences show much lower performances than those in ordinary sentences, and this fact makes it difficult that many systems based on natural language processing (NLP) techniques extract effective lexical features from sentences. To overcome these feature extraction problems for agglutinative languages, we propose a lexical feature extraction method based on longest-match-preference rules of spacing-unit patterns and phoneme patterns. These two patterns are automatically constructed from a large, POS tagged corpus.

2 Feature Extraction Method Based on Longest-Match-Preference Rules

2.1 Feature Extraction Using Spacing-Unit Patterns

Online customer reviews in agglutinative languages frequently include spacing errors. We propose a feature extraction system that matches spacing-unit patterns to input sentences. First, we automatically construct a spacing-unit dictionary from a large, POS tagged corpus. The construction steps are as follows:

1. Extract spaced strings including one or more content words from a POS tagged corpus
2. Remove strings that occur twice or less among the extracted strings
3. Store pairs of key (string itself) and data (content words in the string) in a database

The spacing-unit dictionary consists of key and data pairs: the keys are the surface forms of spaced strings, and the data are stem forms of the keys. In an *eojeol*(Korean spacing unit), functional words follow content words. Therefore, in the Korean spacing-unit dictionary, the keys are the surface forms of *eojeols* and the data are the sequence of content words in the *eojeols*, as shown in Table 1.

Table 1. A section of the Korean spacing-unit dictionary

Key	Data
joh-a-ha-si-ne (be liking)	*joh-a-ha* (like)
joh-ne (be good)	*joh* (good)
ye-ppeo-yo (be pretty)	*i-ppeu* (pretty)

After constructing the spacing-unit dictionary, the proposed system matches an input sentence to each key in the spacing-unit dictionary, using the longest-match-preference rules as follow:

1. Scan an input sentence from left to right
2. Selects the longest match that has a key of the spacing-unit dictionary at each point
3. Return the data associated with the matched key

2.2 Feature Extraction Using Phoneme Patterns

When online users submit customers' reviews, they often input abbreviated or idiomatic words that are intentionally used to reduce the length of sentences or to express social familiarity between online users. For example, the *eojeol* "*joh-a-yo* (like)" is frequently changed into "*jo-a-yo*", "*jo-a-yong*", "*joh-a-yeo*", and so on. To overcome this spelling issue, we propose a feature extraction system that matches phoneme patterns to input sentences. In Korean, intentional spelling errors are closely associated with inner-word abbreviations. Many *eojeols* with similar sounds can be grouped using Kodex [6], a Korean variant of Soundex. To reflect some characteristics of Korean spelling errors (*i.e.*, intentional spelling errors frequently occurring as trailing consonants), we simplify Kodex as follows:

1. Retain all leading consonants and all vowels
2. Drop all trailing consonants
3. Replace some leading consonants with their representative consonants: '*g*', '*gg*' → '*g*' / '*d*', '*dd*' → '*d*' / '*b*', '*bb*' → '*b*' / '*s*', '*ss*' → '*s*' / '*j*', '*jj*' → '*j*' / '*h*', '*p*' → '*p*'

Based on these characteristics of intentional spelling errors in Korean, we can automatically construct a phoneme-unit dictionary from a large POS tagged corpus. In the phoneme-unit dictionary, the keys are a sequence of characters from which the trailing consonants have been removed. The data are content words that are the most frequently occurring among *eojeols* that have been matched to the keys.

Table 2. A section of the Korean phoneme-unit dictionary constructed from *eojeols*

Eojeol		Key	Data
joh-a-ha-si-ne (like)		*jo-a-ha-si-ne*	*joh-a-ha* (like)
joh-a-ha-si-neun (to like)	→	*jo-a-ha-si-neu*	*joh-a-ha* (like)
ye-ppeo-yo (be pretty)		*ye-ppeo-yo*	*ye-ppeu* (pretty)

After constructing the phoneme-unit dictionary, the proposed system matches input sentences to each key in the phoneme-unit dictionary, using the longest-match algorithm mentioned in Section 2.1.

3 Evaluation

3.1 Data Sets and Experimental Settings

To construct a spacing-unit dictionary and a phoneme-unit dictionary, we used the 21st Century Sejong Project's POS tagged corpus (0.8 million sentences, 15 million *Ejoeols*; available at http://www.sejong.or.kr/eindex.php). To evaluate the proposed

feature extraction system experimentally, we collected 12,291 Korean sentences from customers' reviews on an online shopping site (http://shopping.naver.com). 32.64% of sentences (8.15% of total *eojeols*) included spelling variants, and 45.82% of sentences (13.25% of total *eojeols*) included spacing errors. Next, we manually annotated the collected sentences using two sentiment tags, P (positive opinion) and N (negative opinion). We selected 6,235 sentences (4,743 positive sentences and 1,492 negative sentences) that the two annotators had tagged identically. Next, we implemented the sentiment classification system using a LibSVM (set by a default linear kernel function) [7].

3.2 Experimental Results

We implemented two kinds of sentiment classification system. One was a sentiment classification system that used features extracted using the proposed method as SVM (Svpport Vector Machine) inputs. We call this system SentiClass-PM. The other was a sentiment classification system that used features extracted using a conventional morphological analyzer(precision of 95.21%)[8] as SVM inputs. We call this system SentiClass-MA. Table 4 shows the performance of SentiClass-PM compared to that of SentiClass-MA.

Table 3. Effectiveness comparison among feature extraction methods

System	Precision	Recall rate	F1-measure
SentiClass-MA^{GOLD}	0.878	0.787	0.830
SentiClass-MA	0.8345	0.7136	0.7694
SentiClass-PM	0.8451	0.7526	0.7962

In Table 3, Based on this experimental result, we think that the proposed feature extraction method will be more effective than the previous feature extraction methods, which are based on morphological analyzers, in corpus with errors such as customer reviews. In the second experiment, we compared the performance of SentiClass-PM to that of the previous sentiment classification system [9] in the movie review domain (8,000 Korean reviews on the online movie site, http://movie.naver.com), as shown in Table 4. SentiClass-PM outperformed Kim's system (using an optimized syllable kernel function) on all measures, although it used an ordinary linear kernel function based on bag-of-words feature representation as an SVM kernel function. This fact reveals that sentiment classification systems can show good performance if they have effective feature extraction methods, in spite of using conventional classification models.

Table 4. Performance comparison between sentiment classification systems

System	Precision	Recall rate	F1-measure
SentiClass-PM	0.7761	0.7687	0.7724
Kim's system	0.7100	0.7032	0.7066

4 Conclusion

We proposed a feature extraction system for sentiment classification in erroneous sentences such as online customer reviews. First, the system automatically constructs two kinds of pattern dictionaries (the so-called spacing-unit dictionary and the phoneme-unit dictionary) from a large, POS tagged corpus. Next, the system performs longest-matching between input sentences and keys in the pattern dictionaries, and returns content words associated with the longest matched keys. In the experiments on the sentiment classification of erroneous sentences, the proposed system showed better performance than previous systems, which are based on conventional morphological analyzers. On the basis of these experiments, we found that the proposed system is a simple, reliable method of extracting lexical features from erroneous sentences.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A4A01005074). This study was also supported by 2013 Research Grant from Kangwon National University(No. C100 9876-01-01).

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP, pp. 79–86 (2002)
2. Shin, J., Lee, J., Kim, H.: Sentiment categorization of Korean customer reviews using CRFs. In: Proceedings of HCLT, pp. 58–62 (2008) (in Korean)
3. Gretzel, U., Yoo, K.Y.: Use and impact of online travel review. In: Proceedings of the 2008 International Conference on Information and Communication Technologies in Tourism, pp. 35–46 (2008)
4. Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendorf, M., Richards, C.: Normalization of non-standard words. ComputerSpeech and Language 15(3), 287–333 (2001)
5. Xue, Z., Yin, D., Davison, B.D.: Normalizing microtext. In: Proceedings of the AAAI 2011 Workshop on Analyzing Microtext, pp. 74–79 (2011)
6. Kang, B., Lee, J., Choi, K.: Phonetic similarity measure for the Korean transliterations of foreign words. Journal of KIISE: B 26(10), 1143–1259 (1999) (in Korean)
7. Chang, C., Lin, C.: LIBSVM: a library for support vector machines, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
8. Sim, K., Yang, J.: High speed Korean morphological analysis based on adjacency condition check. Journal of KIISE: SA 31(1), 89–99 (2004) (in Korean)
9. Kim, S., Park, S., Park, S., Lee, S., Kim, K.: A syllable kernel based sentiment classification for movie reviews. Journal of KIISS 20(2), 202–207 (2010) (in Korean)

Unified Concept Space and Mapping Discovery Algorithm for Heterogeneous Knowledge Systems*

Lijun Zhu[1], Chen Shi[1], and Jianfeng Guo[2,**]

[1] Institute of Scientific and Technical Information of China, Beijing, China
{zhulj,shichen2012}@istic.ac.cn
[2] Center for Energy & Environmental Policy Research, Institute of Policy and Management, Chinese Academy of Sciences, Beijing, China
guojf@casipm.ac.cn

Abstract. For heterogeneous scientific and technical knowledge systems (HSTKSs), the computer-aided concept-mapping discovery becomes very difficult among the HSTKSs. First, this paper puts forward and establishes a public concept model oriented to the HSTKS used for the standardized description of scientific and technological knowledge concepts. Then, in the mapping discovery algorithm, the algorithms to discover the relations of inheritance, "is a characteristic of", "is a part of", relevance and other partial ordering relations between heterogeneous concepts are put forward and designed through mapping transfer. Finally, the empirical results show that, in public concept space, the mapping discovery algorithm put forward and designed by this paper, is feasible and have certain practical significance.

Keywords: Knowledge System, Concept Space, Mapping Discovery, Mapping Transfer.

1 Introduction

Currently, there are large numbers of classification systems based on different criteria and thesauri of each disciplinary field. Among all the classifications specifications, International Patent Classification (IPC), International Classification for Standards (ICS), Chinese Classification for Standards (CCS), Universal Decimal Classification (UDC), Dewey Decimal Classification (DDC) and Chinese Library Classification (CLC) are widely used. Different to classifications, thesauri are mainly used in China, such as Chinese Classified Thesaurus, Social Science Thesaurus and other general thesauri; and Aerospace Scientific and Technical Thesaurus, Chinese Thesaurus of Petrochemical Industry and so on.

All kinds of scientific and technological knowledge systems need to be reasonably and accurately correlated, in which computer-aided scientific and technological

* This research is granted by National Twelfth "Five-Year Plan" for Science and Technology Support Program: 2011BAH10B04, 2011BAH10B02; National Natural Science Foundation of China 71271200.
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_36,

knowledge concept mapping is a fundamental work practice and an important research direction [1, 2].

Theoretically, if HSTKS is projected to neutral concept space first and then do the research on the method of concept mapping, the computer-aided mapping of heterogeneous knowledge concept will be more effective. For mapping discovery algorithm itself, structured deduction, an important method for it, whose essence is a mapping transfer method, is easy for the mapping deduction algorithm of similarity relation. However, among past researches, few is about mapping transfer algorithm of partial ordering relations, which has largely limited the efficiency of semantic relation discovery between knowledge concepts [3-5].

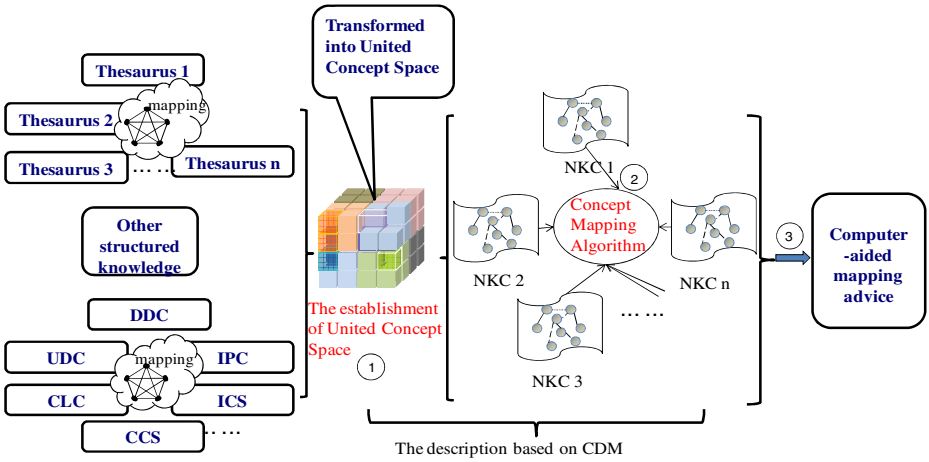


Fig. 1. Research technology roadmap

Research work of this paper includes: First, build a HSTKS oriented to general concept space model, namely Common Data Model (CDM) in the paper, used for standardized description of scientific and technological knowledge; Then, heterogeneous knowledge concepts in CDM will be projected into relevant Neutral Knowledge Concept (NKC), integrating the current semantic analysis technologies, to discuss the computer-aided algorithms of concept mapping among different knowledge systems; At last, verify the efficiency of united concept space model and mapping discovery algorithm by practical calculation. In the research, the research object HSTKS includes thesauri and common classifications, such as IPC, UDC, DDC, CLC, CCS, ICS and so on. The technology roadmap is shown in Fig.1.

2 HSTKS Oriented CDM

2.1 Basic Composition of CDM

Although heterogeneous knowledge systems are different in architecture, there must be some generalities among them. Based on the analysis of all aspects of all classifications and thesauri, the basic constitution of CDM for the HSTKS is summarized as follows: (1) There are 7 kinds of nodes in CDM, which are identified by A to G in

Fig.2. (2) There are 8 kinds of basic semantic relations in CDM, which are identified by L1 to L8 in Fig.2. In Fig.2, the sources of 7 kinds of nodes in CDM are shown on the left part, and 8 kinds of basic semantic relations and their derived potential mapping relations in CDM are shown on the right part.

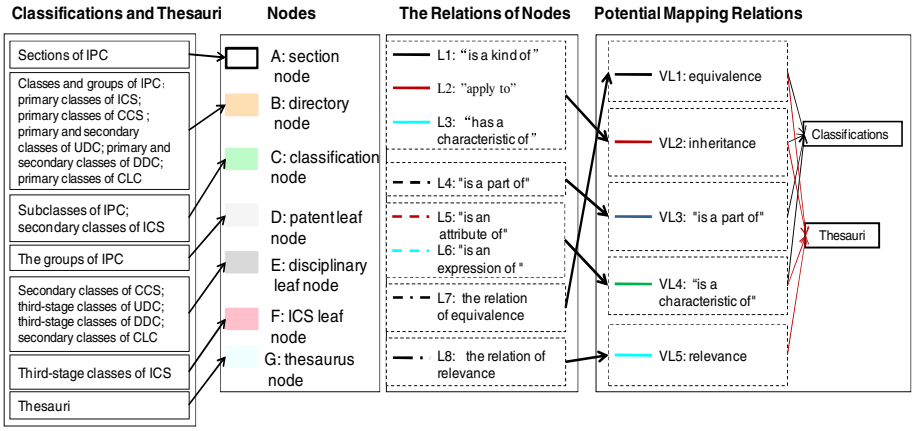


Fig. 2. The nodes and their relations in the architecture

2.2 The Concept Nodes and Their Basic Semantic Relations' Constraints in CDM

In CDM, 7 kinds of concept nodes' coverage are different, so they have different attributes, including 'Name', 'Source', 'Number', 'Description' and 'characteristic'. Between these 7 kinds of nodes, all of the basic semantic relations identified by L1 to L8 are not permitted to exist. The constraints between concept nodes and basic semantic relations are shown in Fig.3.

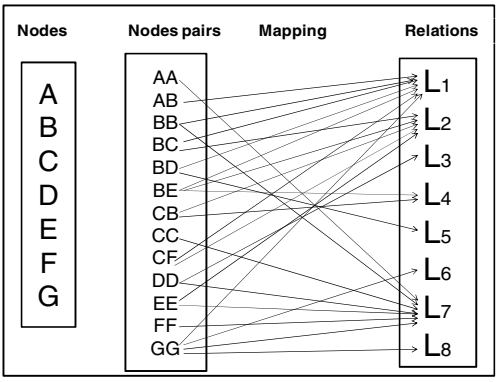


Fig. 3. The relations mapping model between nodes

3 The Concept Mapping Discovery Algorithm Based on CDM

3.1 The Concept Mapping Discovery Algorithm System Based on CDM

On the basis of public data model, this paper designs a set of algorithms of the scientific and technological concept mapping discovery, which can find the specified semantic relations between heterogeneous scientific and technological knowledge concepts. And its architecture is shown in Fig.4.

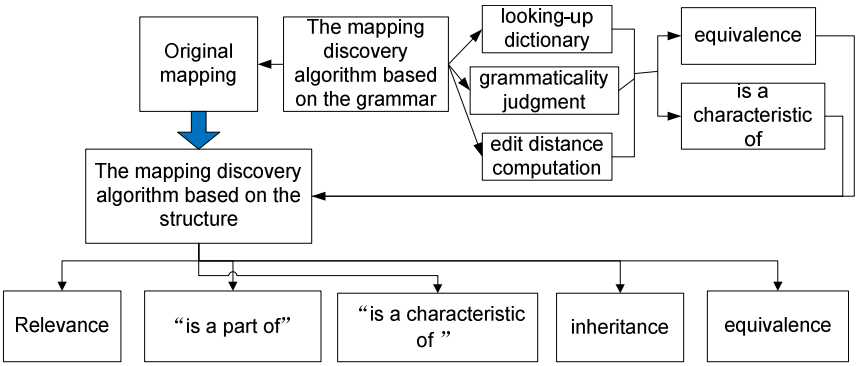


Fig. 4. The architecture of concept mapping discovery algorithm

Using the mature and traditional computing methods, the discovery algorithm of basic relations mapping in grammar level won't be described in detail in the paper.

3.2 The Mapping Discovery Algorithm Based on Structured Deduction

Different to similarity relations, relations of partial order should consider the attenuation effect in the transfer of mapping. In the research, the relations of partial order include 4 kinds: inheritance, "is a characteristic of", "is a part of" and relevance. In data structure, all kinds of classifications and thesauri are tree structures. The transfer toward tree root is called upward transfer, and on the contrary, downward transfer.

3.2.1 The Situation of Upward Transfer

Axiom 1: In the tree knowledge system, no matter mappings of similarity relations or of other partial ordering relations, show no decay in the upward transfer. Namely the transfer decay coefficient is 1 on each step.

3.2.2 The Situation of Downward Transfer

In the tree knowledge system, semantic relations of partial order show a decrease in the downward transfer, however, different partial ordering relations have same algorithm structures and principals. Thus, this paper takes the potential probability

computation of "is a characteristic of" relation for example, which can be seen in the Formula 1.

$$S_{abF} = \sum_{r=1}^{R} \frac{1}{A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}} \alpha_C^{L_1} \alpha_I^{L_2} \alpha_P^{L_3} \alpha_R^{L_4} f_r \quad R = 1,2,\ldots,A_{m+1} + B_{n+1} - A_{m+1} \mathrm{I} B_{n+1} \tag{1}$$

In Formula 1, S_{abF} is the probability computed after the transfer of "is a characteristic of" relation. α_C, α_I, α_P and α_R are transfer decay coefficients between concept nodes to be compared respectively of the four relations of "is a characteristic of", inheritance, "is a part of" and relevance. L_1, L_2, L_3 and L_4 are the numbers of times that the relations of "is a characteristic of", inheritance, "is a part of" and relevance appear. f_r is the similarity of the r[th] attribute between two concepts. A_{m+1} and B_{n+1} are the numbers of attributes belonging to concept A and concept B. $A_{m+1} \cap B_{n+1}$ is the number of the attributes of same type in concept A and concept B. If only the attributes of English names are compared in concept nodes, the formula 1 can be simplified as the following formula 2.

$$S_{abF} = \alpha_C^{L_1} \alpha_I^{L_2} \alpha_P^{L_3} \alpha_R^{L_4} f_{EN} \tag{2}$$

Generally, in the same scientific and technological systems, the probability that some semantic relation appears has uniform distribution, namely, independent of the sample size. So suppose that the decay coefficients α_C, α_I, α_P and α_R should be relevant to the average value $\overline{P}$ of $\overline{P_A}$ and $\overline{P_B}$ which are the proportions that the corresponding semantic relations account for all relations respectively in two classifications(also a classification and a thesaurus or two thesauri), and relevant to the probability of equivalence S between the transfer's original two nodes. Moreover, the decay coefficients α_C, α_I, α_P and α_R need to have some constraints: First, α_C, α_I, α_P and α_R must be monotone functions; And then, if S is 0, α_C, α_I, α_P and α_R must be 0 and if S is 1, α_C, α_I, α_P and α_R must be 1; Finally, the range of S is from 0 to 1. Therefore, the calculation functions of α_C, α_I, α_P and α_R are defined as $\alpha = \frac{(\overline{P}+1)S}{\overline{P}+S}$, with α_C, α_I, α_P and α_R corresponding respectively to $\overline{P}_C$, $\overline{P}_I$, $\overline{P}_P$ and $\overline{P}_R$.

4 The Empirical Results

Experimental samples from Petroleum Thesaurus, IPC and UDC of petroleum-related area are described by CDM. The number of concepts used for experiment extracted from Petroleum Thesaurus is 1500, and 900 concepts possibly related to petroleum are extracted from UPC, as well as 700 from UDC.

Table 1. The results of the mapping relations found and verified artificially

Sources		Total	E-R	I-R	R-R	IACO-R
PT&IPC	P	231	161	42	28	0
	V	\	11	17	26	\
PT&UDC	P	238	105	112	21	0
	V	\	7	37	19	\
IPC&UDC	P	161	29	97	28	7
	V	\	7	16	23	2

Note: P is the number of potential mapping relations found and V is the number of mapping relations that are verified artificially. PT&IPC is the number of the mapping relations found between Petroleum Thesaurus and IPC. PT&UDC is the number of the mapping relations found between Petroleum Thesaurus and UDC. IPC&UDC is the number of the mapping relations found between IPC and UDC. E-R is equivalence relation. I-R is inheritance relation. R-R is relevance relation. IACO-R is “is a characteristic of “relation.

The experimental results: (1) Totally, 231 potential mapping relations are found between Petroleum Thesaurus and IPC (PT&IPC): The number of the potential equivalence relations is 161 and the number verified artificially is 11; The number of the potential inheritance relations is 42 and the number verified artificially is 17; The number of the potential relevance relations is 28 and the number verified artificially is 26. (2) Totally, 238 potential mapping relations are found between Petroleum Thesaurus and UDC (PT&UDC): The number of the potential equivalence relations is 105 and the number verified artificially is 7; The number of the potential inheritance relations is 112 and the number verified artificially is 37; The number of the potential relevance relations is 21 and the number verified artificially is 19. (3) Totally, 161 potential mapping relations are found between IPC and UDC (IPC&UDC): The number of the potential equivalence relations is 29 and the number verified artificially is 7; The number of the potential inheritance relations is 97 and the number verified artificially is 16; The number of the potential relevance relations is 28 and the number verified artificially is 23; The number of the potential "is a characteristic of " relations is 7 and the number verified artificially is 2.

5 The Conclusion, Deficiency and Prospect

The CDM put forward and established by this paper could projects HSTKS to the same concept space, which making computer-aided discovery possible among various HSTKSs. The algorithm to discover the partial ordering semantic relations is proposed and designed based on mapping transfer, providing a new research thought for the computer-aided discovery of the defined semantic relations. The experimental results indicate that the algorithm has certain practical value.

With the algorithm experiment done between the thesaurus, IPC and UDC, the potential semantic mapping of the "is a part of" relation is not found in the experimental results, and the percent of the effective mapping is not high. So the accuracy of the projection from HSTKS to CDM should be improved, and the design of CDM could be further analyzed and optimized.

References

1. Adali, S., Emery, R.: A uniform framework for integrating knowledge in heterogeneous knowledge systems. In: Proceedings of the Eleventh International Conference on IEEE, Data Engineering, pp. 513–520 (March 1995)
2. Gruber, T.R.: A translation approach to portable ontology specifications. Technical Report of Knowledge System Laboratory, 513–520 (1993)
3. Farquhar, A., Fikes, R., Rice, J.: The Ontolingua server: A tool for collaborative ontology construction. Int'l Journal of Human-Computer Studies 46(6), 707–727 (1997)
4. Czarnecki, K., Antkiewicz, M.: Mapping features to models: A template approach based on superimposed variants. In: Glück, R., Lowry, M. (eds.) GPCE 2005. LNCS, vol. 3676, pp. 422–437. Springer, Heidelberg (2005)
5. Gruber, T.: Collective knowledge systems: Where the social web meets the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 6(1), 4–13 (2008)

Author-Topic over Time (AToT): A Dynamic Users' Interest Model

Shuo Xu[1], Qingwei Shi[1,2], Xiaodong Qiao[3,⋆], Lijun Zhu[1], Hanmin Jung[4], Seungwoo Lee[4], and Sung-Pil Choi[4]

[1] Information Technology Supporting Center, Institute of Scientific and Technical Information of China, No. 15 fuxing Rd., Haidian District, Beijing 100038, P.R. China
[2] School of Software, Liaoning Technical University, No. 188 Longwan St. South, Huludao, Liaoning 125105, P.R. China
[3] Colledge of Software, Northeast Normal University, 5268 Renmin St., Changchun, Jilin 130024, P.R. China
[4] Department of Computer Intelligence Research, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea
{xush,shiqw,qiaox,zhulj}@istic.ac.cn, {jhm,swlee,spchoi}@kisti.re.kr

Abstract. One of the key problems in upgrading information services towards knowledge services is to automatically mine latent topics, users' interests and their evolution patterns from large-scale S&T literatures. Most of current methods are devoted to either discover static latent topics and users' interests, or to analyze topic evolution only from intra-features of documents, namely text content without considering directly extra-features of documents such as authors. To overcome this problem, a dynamic users' interest model for documents using authors and topics with timestamps is proposed, named as Author-Topic over Time (AToT) model, and collapsed Gibbs sampling method is utilized for inferring model parameters. This model is not only able to discover latent topics and users' interests, but also to mine their changing patterns over time. Finally, the extensive experimental results on NIPS dataset with 1,740 papers indicate that our AToT model is feasible and efficient.

Keywords: Author-Topic (AT) Model, Topic over Time (ToT) Model, Author-Topic over Time (AToT) Model, Dynamic Users' Interest Model, Collapsed Gibbs Sampling.

1 Introduction

With a dynamic users' interest model, one can answer a range of important questions about the content of document collections, such as which topics each user prefers to, which users are similar to each other in terms of their interests, which users are likely to have written documents similar to an observed document, and

⋆ Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_37,

who are influential users at different stages of topic evolution and it also helps characterize users as pioneers, mainstream or laggards in different subject areas. Users' interests have shown their increasing importance for the development of personalized Web services and user-centric applications [1,2]. Hence, users' interest modeling has been attracting extensive attentions during the past few years, such as (a) Author-Topic (AT) model [3]; (b) Author-Recipient-Topic (ART) [4], Role-Author-Recipient-Topic (RART) [4] & Author-Persona-Topic (APT) models [5]; (c) Author-Interest-Topic (AIT) [6] & Latent-Interest-Topic (LIT) models [7], and (d) Author-Conference-Topic (ACT) model [8], etc.

In fact, in the process of entire scientific career, each researcher's interest is usually not static. However, the above models are devoted to discover static latent topics and research interests. Of course, one can perform some post-hoc or pre-hoc analysis [9,10] to discover changing patterns over time, but this misses the opportunity for time to improve topic discovery [11], and it is very difficult to align corresponding topics [12].Currently, attention for dynamic models is mainly focused on analyzing topic evolution only from text content, such as Dynamic Topic Model (DTM) [13], continuous time DTM (cDTM) [14], Topic over Time (ToT) [11], and so on.

This article mainly focuses on the dynamic users' interest model. The organization of the rest of this paper is as follows. In Sec. 2, we discuss generative models for documents using authors and topics with timestamps, introduce the Author-Topic over Time (AToT) model in detail on the basis of AT and ToT models and describe the collapse Gibbs sampling methods used for inferring the model parameters. In Sec. 3, extensive experimental evaluations are conducted, and Sec. 4 concludes this work.

2 Author-Topic over Time (AToT) Model

The notation is summarized in Table 1, and the graphical model representations of the AToT model is shown in Fig. 1. The AToT model can be viewed as a generative process, which can be described as follows.

Table 1. Notation used in the AToT model

SYMBOL	DESCRIPTION
K	Number of topics
M	Number of documents
V	Number of unique words
A	Number of unique authors
N_m	Number of word tokens in document m
A_m	Number of authors in document m
$\boldsymbol{a}_m$	Authors in document m
$\boldsymbol{\vartheta}_a$	Multinomial distribution of topics specific to the author a. And let $\boldsymbol{\Theta} = \{\boldsymbol{\vartheta}_a\}_{a=1}^{A}$
$\boldsymbol{\varphi}_k$	Multinomial distribution of words specific to the topic k. And let $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_k\}_{k=1}^{K}$
$\boldsymbol{\psi}_k$	Beta distribution of timestamp specific to the topic k. And let $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_k\}_{k=1}^{K}$
$z_{m,n}$	Topic associated with the n-th token in the document m
$w_{m,n}$	n-th token in document m
$x_{m,n}$	Chosen author associated with the word token $w_{m,n}$
$t_{m,n}$	Timestamp associated with the n-th token in the document m
$\boldsymbol{\alpha}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\boldsymbol{\vartheta}$
$\boldsymbol{\beta}$	Dirichlet priors (hyperparameter) to the multinomial distribution $\boldsymbol{\varphi}$

1. For each topic $k \in [1, K]$ and each author $a \in [1, A]$, draw a $\boldsymbol{\varphi}_k \sim \text{Dirichlet}(\boldsymbol{\beta})$ and $\boldsymbol{\vartheta}_a \sim \text{Dirichlet}(\boldsymbol{\alpha})$, respectively;
2. For each word $n \in [1, N_m]$ in document $m \in [1, M]$:
   - Draw an author assignment $x_{m,n} \sim \text{Uniform}(\boldsymbol{a}_m)$;
   - Draw a topic assignment $z_{m,n} \sim \text{Multinomial}(\boldsymbol{\vartheta}_{x_{m,n}})$;
   - Draw a word $w_{m,n} \sim \text{Multinomial}(\boldsymbol{\varphi}_{z_{m,n}})$;
   - Draw a timestamp $t_{m,n} \sim \text{Beta}(\psi_{z_{m,n},1}, \psi_{z_{m,n},2})$;

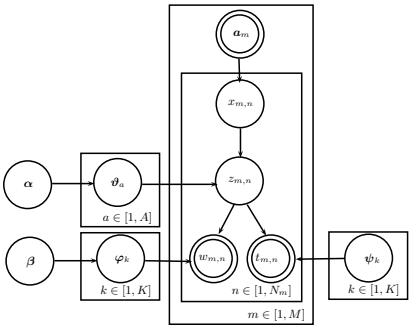


Fig. 1. The graphical model representation of the AToT model

For inference, the task is to estimate the sets of following unknown parameters in the AToT model: (1) $\boldsymbol{\Phi}, \boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$; (2) the corresponding topic and author assignments $z_{m,n}, x_{m,n}$ for each word token $w_{m,n}$. In fact, inference can not be done exactly in this model. In this work, collapsed Gibbs sampling algorithm [15] is used, since it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows combination of estimates from several local maxima of the posterior distribution.

In the Gibbs sampling procedure, we need to calculate the conditional distribution $P(z_{m,n}, x_{m,n} | \boldsymbol{w}, \boldsymbol{z}_{\neg(m,n)}, \boldsymbol{x}_{\neg(m,n)}, \boldsymbol{t}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi})$ with $\boldsymbol{z}_{\neg(m,n)}$, $\boldsymbol{x}_{\neg(m,n)}$ represents the topic, author assignments for all tokens except $w_{m,n}$, respectively. We begin with the joint distribution $P(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{t} | \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi})$ of a dataset, and using the chain rule, we can get the conditional probability conveniently as

$$
\begin{aligned}
&P(z_{m,n} = k, x_{m,n} = a | \boldsymbol{w}, \boldsymbol{z}_{\neg(m,n)}, \boldsymbol{x}_{\neg(m,n)}, \boldsymbol{t}, \boldsymbol{a}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}) \\
&\propto \frac{n_k^{(w_{m,n})} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^{V} (n_k^{(v)} + \beta_v) - 1} \times \frac{n_a^{(k)} + \alpha_k - 1}{\sum_{k=1}^{K} (n_a^{(k)} + \alpha_k) - 1} \times \text{Beta}(\psi_{z_{m,n},1}, \psi_{z_{m,n},2}) \quad (1)
\end{aligned}
$$

where $n_k^{(v)}$ is the number of times tokens of word v is assigned to topic k, and $n_a^{(k)}$ represents the number of times author a is assigned to topic k.

During parameter estimation, the algorithm keeps track of two large data structures: an $A \times K$ count matrix $n_a^{(k)}$ and an $K \times V$ count matrix $n_k^{(v)}$.

From these data structures, one can easily estimate the $\mathbf{\Phi}$ and $\mathbf{\Theta}$ as follows: $\varphi_{k,v} = \frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)}$ and $\vartheta_{a,k} = \frac{n_a^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_a^{(k)}+\alpha_k)}$. As for $\mathbf{\Psi}$, for simplicity and speed we update it after each Gibbs sample by the method of moments: $\psi_{k,1} = \bar{t}_k\left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2}-1\right)$ and $\psi_{k,2} = (1-\bar{t}_k)\left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2}-1\right)$, where $\bar{t}_k$ and s_k^2 indicate the sample mean and biased sample variance of the timestamps belonging to topic k, respectively. The readers are invited to consult [16] for details. Note that the time range of the data is normalized to [0.01, 0.99].

3 Experimental Results and Discussions

NIPS proceeding dataset is utilized to evaluate the performance of our model, which consists of the full text of the 13 years of proceedings from 1987 to 1999 Neural Information Processing Systems (NIPS) Conferences. In addition to downcasing and removing stopwords and numbers, we also removed the words appearing less than five times in the corpus. The dataset contains 1,740 research papers, 2,037 unique authors, 13,649 unique words, and 2,301,375 word tokens in total. Each document's timestamp is determined by the year of the proceedings. In our experiments, K is fixed at 100, and the symmetric Dirichlet priors α and β are set at 0.5 and 0.1, respectively. Gibbs sampling is run for 2000 iterations.

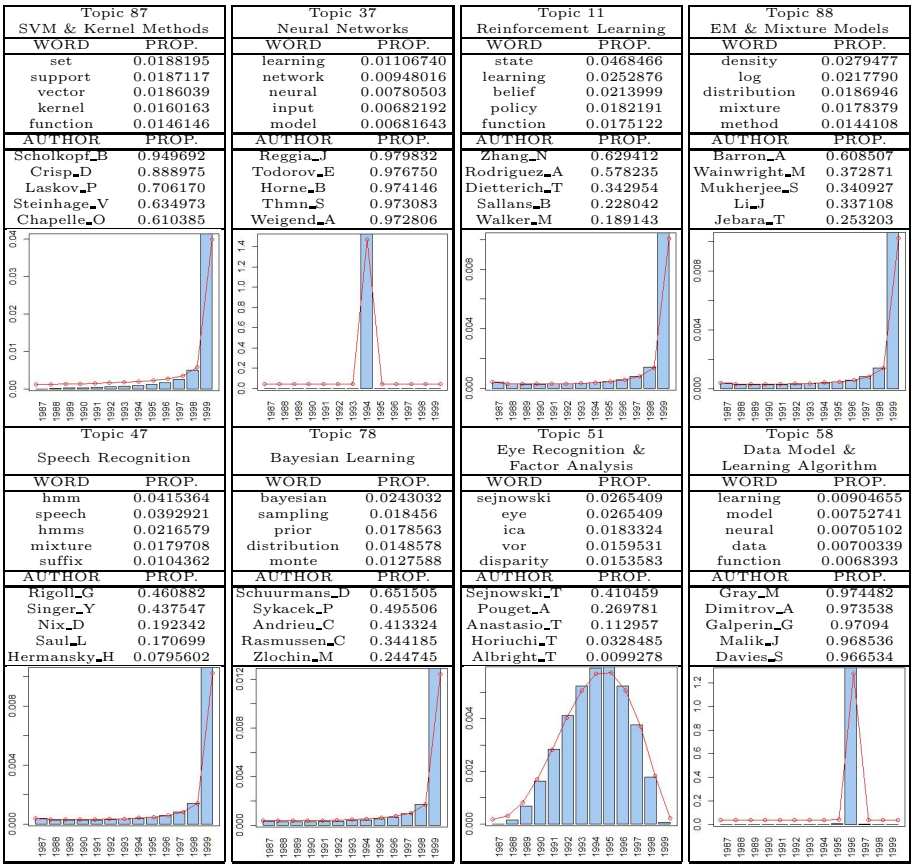
3.1 Examples of Topic, Author Distributions and Topic Evolution

Fig. 2 illustrates examples of 8 topics learned by AToT model. The topics are extracted from a single sample at the 2000th iteration of the Gibbs sampler. Each topic is illustrated with (1) the top 5 words most likely to be generated conditioned on the topic; (b) the top 5 authors which have the highest probability conditioned on the topic; and (c) histograms and fitted beta PDFs which show topics evolution patterns over time.

3.2 Author Interest Evolution Analysis

In order to analyze further author interest evolution, it is interesting to calculate $P(z,t|a) = P(z|a)p(z|t) = \boldsymbol{\vartheta}_{a,z} \times \text{Beta}(\psi_{z,1}, \psi_{z,2})$. In this subsection, we take Sejnowski as an example, who published 43 papers in total from 1987 to 1999 in the NIPS conferences, as shown Fig. 3 (a). The research interest evolution for Sejnowski is reported in Fig. 3 (b), in which the area occupied by a square is proportional to the strength of his research interest.

From Fig. 3 (b), one can see that Sejnowski's research interest focused mainly on Topic 51 (Eye Recognition & Factor Analysis), Topic 37 (Neural Networks) and Topic 58 (Data Model & Learning Algorithm) but with different emphasis from 1987 to 1999. In the early phase (1989–1993), Sejnowski's research interest is only limited to Topic 51, and then extended to Topic 37 in 1994 & Topic 58 in 1996 with great research interest strength, and finally back to Topic 51 after 1997. Anyway, Sejnowski did not change his main research direction, Topic 51, which is verified from his homepage again.



Topic 87 SVM & Kernel Methods	
WORD	PROP.
set	0.0188195
support	0.0187117
vector	0.0186039
kernel	0.0160163
function	0.0146146
AUTHOR	PROP.
Scholkopf_B	0.949692
Crisp_D	0.888975
Laskov_P	0.706170
Steinhage_V	0.634973
Chapelle_O	0.610385

Topic 37 Neural Networks	
WORD	PROP.
learning	0.01106740
network	0.00948016
neural	0.00780503
input	0.00682192
model	0.00681643
AUTHOR	PROP.
Reggia_J	0.979832
Todorov_E	0.976750
Horne_B	0.974146
Thmn_S	0.973083
Weigend_A	0.972806

Topic 11 Reinforcement Learning	
WORD	PROP.
state	0.0468466
learning	0.0252876
belief	0.0213999
policy	0.0182191
function	0.0175122
AUTHOR	PROP.
Zhang_N	0.629412
Rodriguez_A	0.578235
Dietterich_T	0.342954
Sallans_B	0.228042
Walker_M	0.189143

Topic 88 EM & Mixture Models	
WORD	PROP.
density	0.0279477
log	0.0217790
distribution	0.0186946
mixture	0.0178379
method	0.0144108
AUTHOR	PROP.
Barron_A	0.608507
Wainwright_M	0.372871
Mukherjee_S	0.340927
Li_J	0.337108
Jebara_T	0.253203

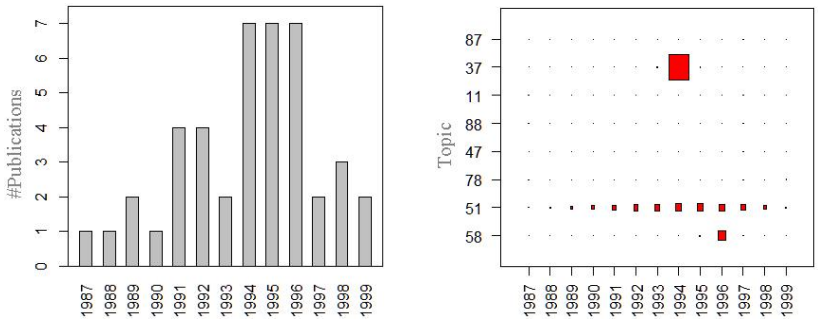
Topic 47 Speech Recognition	
WORD	PROP.
hmm	0.0415364
speech	0.0392921
hmms	0.0216579
mixture	0.0179708
suffix	0.0104362
AUTHOR	PROP.
Rigoll_G	0.460882
Singer_Y	0.437547
Nix_D	0.192342
Saul_L	0.170699
Hermansky_H	0.0795602

Topic 78 Bayesian Learning	
WORD	PROP.
bayesian	0.0243032
sampling	0.018456
prior	0.0178563
distribution	0.0148578
monte	0.0127588
AUTHOR	PROP.
Schuurmans_D	0.651505
Sykacek_P	0.495506
Andrieu_C	0.413324
Rasmussen_C	0.344185
Zlochin_M	0.244745

Topic 51 Eye Recognition & Factor Analysis	
WORD	PROP.
sejnowski	0.0265409
eye	0.0265409
ica	0.0183324
vor	0.0159531
disparity	0.0153583
AUTHOR	PROP.
Sejnowski_T	0.410459
Pouget_A	0.269781
Anastasio_T	0.112957
Horiuchi_T	0.0328485
Albright_T	0.0099278

Topic 58 Data Model & Learning Algorithm	
WORD	PROP.
learning	0.00904655
model	0.00752741
neural	0.00705102
data	0.00700339
function	0.0068393
AUTHOR	PROP.
Gray_M	0.974482
Dimitrov_A	0.973538
Galperin_G	0.97094
Malik_J	0.968536
Davies_S	0.966534

Fig. 2. An illustration of 8 topics from a 100-topic solutions for the NIPS collection. The titles are our own interpretation of the topics. Each topic is shown with the 5 words and authors that have the highest probability conditioned on that topic. Histograms show how the topics are distributed over time; the fitted beta PDFs is shown also.



(a) Distribution of #publications over time (b) Research Interest Evolution

Fig. 3. The distribution of #publications and research interest evolution for Sejnowski

3.3 Predictive Power Analysis

Similar to [3], we further divide the NIPS papers into a training set $\mathcal{D}^{\text{train}}$ of 1,557 papers, and a test set $\mathcal{D}^{\text{test}}$ of 183 papers of which 102 are single-authored papers. Each author in $\mathcal{D}^{\text{test}}$ must have authored at least one of the training papers. The perplexity is a standard measure for estimating the performance of a probabilistic model. The perplexity of a test document $\tilde{m} \in \mathcal{D}^{\text{test}}$, is defined as the exponential of the negative normalized predictive likelihood under the model: $\text{perplexity}(\boldsymbol{w}_{\tilde{m},\cdot}, \boldsymbol{t}_{\tilde{m},\cdot}|\boldsymbol{a}_{\tilde{m}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \exp\left[-\frac{\ln P(\boldsymbol{w}_{\tilde{m},\cdot}, \boldsymbol{t}_{\tilde{m},\cdot}|\boldsymbol{a}_{\tilde{m}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi})}{N_{\tilde{m}}}\right]$ with

$$P(\boldsymbol{w}_{\tilde{m},\cdot}, \boldsymbol{t}_{\tilde{m},\cdot}|\boldsymbol{a}_{\tilde{m}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \frac{1}{[A_{\tilde{m}}]^{N_m}} \times \sum_{\boldsymbol{z}_{\tilde{m},\cdot}} \text{Beta}(\psi_{z_{\tilde{m},n},1}, \psi_{z_{\tilde{m},n},2}|\mathcal{D}^{\text{train}}) \times$$

$$\int p(\boldsymbol{\Phi}|\boldsymbol{\beta}, \mathcal{D}^{\text{train}}) \sum_{\boldsymbol{z}_{\tilde{m},\cdot}} \varphi_{z_{\tilde{m},n}, w_{\tilde{m},n}} d\boldsymbol{\Phi} \times \int p(\boldsymbol{\Theta}|\boldsymbol{\alpha}, \mathcal{D}^{train}) \sum_{\boldsymbol{x}_{\tilde{m},\cdot}} \vartheta_{x_{\tilde{m},n}, z_{\tilde{m},n}} d\boldsymbol{\Theta} \quad (2)$$

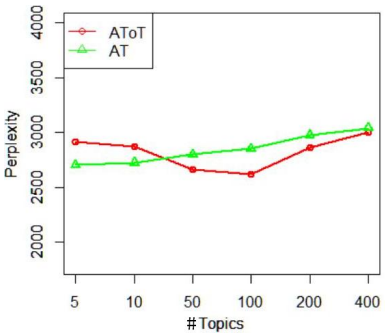


Fig. 4. Perplexity of the 102 single-authored test documents

We approximate the integrals over $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ using the point estimates obtained in Sec. 2 for each sample $s \in \{1, 2, \cdots, 10\}$ of assignments $\boldsymbol{x}, \boldsymbol{z}$, and then average over samples. Fig. 4 shows the results for the AToT model and AT model in a post-hoc fashion on 102 single-authored papers. It is not difficult to see that the perplexity of AToT model is smaller than that of AT model when #topics > 10, which indicates that AToT model outperforms AT model.

4 Conclusions

With a dynamic users' interest model, one can answer many important questions about the content of document collections. Based on AT & ToT models, this article proposes a dynamic users' interest model, Author-Topic over Time (AToT) model, for documents using authors and topics with timestamps, and collapsed Gibbs sampling is used for inferring model parameters. It combines the

merits of AT & ToT models. The results on NIPS dataset show the discovery of more salient topics and more reasonable users' interest evolution patterns. What's more, one can generalize the approach in the work to construct alternative dynamic models from other static users' interest models and ToT model.

Acknowledgments. This work was funded partially by Key Technologies R&D Program of Chinese 12th Five-Year Plan (2011–2015): Key Technologies Research on Large-Scale Semantic Calculation for Foreign STKOS, and Key Technologies Research on Data Mining from the Multiple Electric Vehicle Information Sources under grant number 2011BAH10B04 and 2013BAG06B01, respectively.

References

1. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: WWW 2006, pp. 727–736. ACM, New York (2006)
2. Kim, J., Jeong, D.H., Lee, D., Jung, H.: User-centered innovative technology analysis and prediction application in mobile environment. Multimed. Tools Appl. (2013)
3. Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., Steyvers, M.: Learning author-topic models from text corpora. ACM T. Inform. Syst. 28(1), 1–38 (2010)
4. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in socail networks with experiments on enron and academic email. J. Artif. Intell. Res. 30(1), 249–272 (2007)
5. Mimno, D., McCallum, A.: Expertise modeling for matching papers with reviewers. In: KDD 2007, pp. 500–509. ACM, New York (2007)
6. Kawamae, N.: Author interest topic model. In: SIGIR 2010, pp. 887–888. ACM, New York (2010)
7. Kawamae, N.: Latent interest-topic model: Finding the causal relationships behind dyadic data. In: CIKM 2010, pp. 649–658. ACM, New York (2010)
8. Tang, J., Zhang, J., Jin, R., Yang, Z., Cai, K., Zhang, L., Su, Z.: Topic level expertise search over heterogeneous networks. Mach. Learn. 82(2), 211–237 (2011)
9. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: KDD 2004, pp. 306–315. ACM, New York (2004)
10. Wang, X., Mohanty, N., McCallum, A.: Group and topic discovery from relations and their attributes. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) NIPS18, pp. 1449–1456. MIT Press, Cambridge (2006)
11. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: KDD 2006, pp. 424–433. ACM, New York (2006)
12. Xu, S., Zhu, L., Qiao, X., Shi, Q., Gui, J.: Topic linkages between papers and patents. In: AST 2012. SERSC, pp. 176–183. Daejeon, South Korea (2012)
13. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML 2006, pp. 113–120. ACM, New York (2006)
14. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: UAI 2008, pp. 579–586 (2008)
15. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. USA 101(suppl. 1), 5228–5235 (2004)
16. Owen, C.B.: Parameter estimation for the Beta distribution. Master's thesis, Brigham Young University (2008)

Scalable RDF Path Query Processing Based on Runtime Class Path Lookup Scheme

Sung-Jae Jung[1,2], Dong-min Seo[1], Seungwoo Lee[1], and Hanmin Jung[1,2]

[1] Department of Computer Intelligence Research,
Korea Institute of Science and Technology Information (KISTI),
245 Daehak-ro, Yuseong-gu, Daejeon, 305-806, Korea
[2] Department of Knowledge and Inforamtion Science,
University of Science and Technology (UST),
217 Gajeong-ro, Yuseong-gu, Daejeon, 305-350, Korea
{sjjung,dmseo,swlee,jhm}@kisti.re.kr

Abstract. With the rapidly growing amount of information represented in RDF format, efficient querying RDF graph has become a fundamental challenge. There have been several relationship finding services based on querying RDF database to discover relationships between two objects of interest. Conventional relationship-finding service requires computationally expensive graph search operations which involve multiple self joins. It becomes even more challenging when the graph data is large and diverse. In this paper we propose an algorithm which uses RDF schema information for efficient RDF path query processing. By utilizing the pre-calculated class path expressions, the graph search space is significantly reduced. Compared with the conventional BFS algorithm, the proposed algorithm (bidirectional BFS combined with class path lookup approach) achieves performance improvement by 3 orders of magnitude. Additionally, the proposed algorithm is scalable, because it operates based on B-Tree index when it accesses to triple repository and pre-calculated class path information. Thus, the proposed algorithm is expected to return graph search results within a reasonable response time on even much larger RDF graph.

Keywords: RDF schema, path expression, SQL based graph search, RDF path query, class path pre-calculation, bidirectional Breadth First Search.

1 Introduction

Since the Semantic Web emerged as a vision of the next generation World Wide Web, the amount of information represented in RDF format has been rapidly growing [1]. Based on the richness of RDF dataset, several relationship finding services have been proposed. Microsoft co-author path[1], Relfinder[2] and OntoRelfinder[3] are the examples of relationship finding services which retrieve relationships between two given objects of interest from RDF data and display the result as a graph.

[1] http://academic.research.microsoft.com/VisualExplorer

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_38,

These services require computationally expensive graph search operations which involve multiple self joins. Abadi et al. [4] materialized path expressions to reduce self joins in path query processing. They pre-calculated the selected path expressions and stored the result in relational tables. OntoRelfinder [3] also pre-calculated and materialized path expressions to reduce search space but the scope of them remained within the triples of the RDF schema. They call them ClassPaths. However, in order to process a user query, OntoRelfinder generates a set of SPARQL query by using the pre-calculated ClassPaths. In case the RDF schema is not simple enough, OntoRelfinder might generate more SPARQL queries than it can handle within a reasonable response time. In this study we adopt the way OntoRelfinder pre-calculates and materializes class paths but we propose a different path querying scheme that utilizes the pre-calculated class paths.

In this paper we propose a novel RDF path querying scheme which utilizes bidirectional BFS algorithm combined with class path lookup scheme to reduce search space in path query processing. The rest of this paper is organized as follows. In section 2, we describe our approach to reduce search space by utilizing the class path information. In section 3, we describe the results of the performance test of the proposed approach. In section 4, we describe the conclusion and the prospected future works.

2 Path Query Processing Based on RDF Schema Information

The proposed approach is a kind of heuristic graph search algorithm that can significantly save the search cost by using ontology schema information. In order to reduce graph search space by using the materialized class path, we propose an algorithm that is modified from Breadth First Search (BFS) algorithm, namely BFS combined with runtime Class Path Lookup scheme.

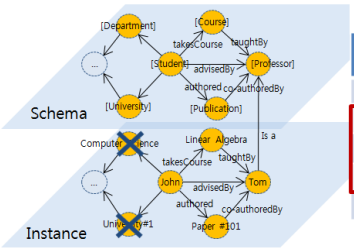
2.1 Class Path Lookup Operation

The main modification is adding a filter operation to the BFS algorithm, which we call ‘Class Path Lookup Operation’. The operation filters out the non-promising neighbor nodes by looking up the pre-calculated class path expressions stored in a relational table. Fig.1 shows the structure of the relational table for class path expressions. Since the filtering operation leverages the B-tree index composed of multiple columns (namely *Start Class, End Class, Path Length and Class Path*), it is executed with the minimal cost.

2.2 Combining BFS with Class Path Lookup Operation

Our algorithm first searches the neighbors of the starting node as BFS does; then, looks up the materialized path expressions to see which neighbor node can lead to the destination successfully within the given path length. The algorithm enqueues only the promising nodes and discards the nodes which lead to unsuccessful search. The neighbor node search operation also utilizes B-Tree index. Fig.2 also shows the way

the proposed algorithm processes an example path query, 'Find relationships between John who is a student and Tom who is a professor within path length 2'. In this example, 'John' has 5 neighbor nodes which include 'Linear Algebra', 'Paper #101', 'Computer Science', and so on. The algorithm enqueues 'Linear Algebra' node for the next search step because the 'Class Path Lookup operation' finds the class path that starts with 'Student/Course' and ends in 'Professor'. On the other hand, the algorithm discards 'Computer Science', which is a 'Department', because the 'Class Path Lookup Operation' fails to find any class path that starts with 'Student/ Department' and ends in 'Professor'.



Materialized class path expressions

start class	end class	Path length	Class path	Predicate path
...	...	...	...	...
student	professor	1	student/professor	advisor
student	professor	2	student/course/professor	takesCourse/taughtBy
student	professor	2	student/publication/professor	authored/co-authoredBy
...	...	...	...	...

Fig. 1. Example of how BFS combined with Class Path Lookup operation works

Fig. 2 describes the proposed algorithm in pseudo code. The algorithm is derived by modifying the conventional BFS algorithm. Line 14 is the filtering operation by class path lookup, where the class path of each neighbor node denoted by *u.ClassPath* in the pseudo code is looked up against the materialized class path table to see if there exists any class path which starts with *u.ClassPath* and ends in the destination class D. For this operation, each node should remember classes of the previously traversed nodes, which is done in line 3 and 13.

```
Input  G : a graph G,
       s, S, d, D : start instance, start Class, destination instance and destination Class in G respectively
       P : a set of class paths pre-calculated from RDF Schema graph that start from S and end in D
1  procedure BFS combined with class path lookup
2      create a queue Q
3      s.ClassPath ←S; s.InstancePath ←s;
4      enqueue s onto Q
5      mark s
6      while Q is not empty
7          t ← Q.dequeue
8          if t equals d :
9              return t
10         for all edges e in G.adjacentEdges(t) do
11             u ← G.adjacentVertex(t,e)
12             if u is not visited:
13                 u.ClassPath ←t.ClassPath+u.ClassPath
14                 if exists any class path in P that starts with u.ClassPath
15                     u.InstancePath ←t.InstancePath+u.InstancePath
16                     enqueue u onto Q
17         end for
```

Fig. 2. Pseudo Code of BFS combined with Class Path Lookup Scheme

2.3 Combining Bidirectional BFS with Class Path Lookup Operation

The 'Class Path Lookup Operation' can also be combined with bidirectional BFS algorithm, which further reduces the search space. The bidirectional BFS initiates its search from two nodes (source and destination nodes) within half the given path length. By joining the two result sets, one expanded from source and the other one expanded from destination, the set of entire paths are returned as a final result. For each step the bidirectional BFS searches neighbors, 'Class Path Lookup Operation' is executed to filter out the non-promising nodes which cannot lead to destination node in forward search or which cannot lead to source node in backward search.

3 Implementation and Performance Evaluation

All the proposed algorithms are implemented by using SQL. The recursive SQL query is written based on common table expressions(CTEs) which are the standard SQL syntax [5] supported by IBM DB2, MS SQL Server, Oracle and so on. The initial search result set retrieved from the triple table is stored in user memory buffer, which is then self-joined to triple table in recursive manners. The schematic of recursive CTE query that executes BFS search is represented in Fig.4, where T1 denotes triple table.

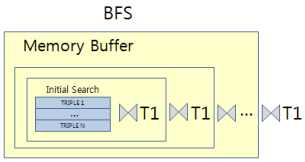


Fig. 3. Schematic of BFS based on Recursive Self Join

We run our experiments on a 1.7 GHz quad core processor with 4GB main memory running Window 7 with the LUBM (10,0) dataset [6] which has about 1.3 million triples. The LUBM data set is synthetically generated by the data generator UBA 1.7[2] and converted into n-triple format, which is then loaded onto a relational table. We used Oracle 11gR2 as a triple repository with 2 GB main memory assigned.

Fig.4 shows the performance of our approach. As can be seen in the figure, combining the 'Filter by Class Path Lookup' with BFS significantly improves the path query processing performance. The performance test was conducted varying the graph search path length. As the path length increases, the performance difference between BFS and the proposed approach grows exponentially. In case of path length 6, our approach (bidirectional BFS combined with 'Class Path Lookup Operation') achieves performance improvement by more than 3 orders of magnitude.

[2] http://swat.cse.lehigh.edu/projects/lubm/

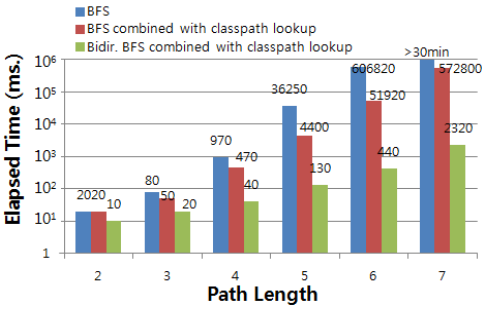


Fig. 4. Performance of the Proposed Algorithm

4 Conclusion

We proposed a novel path querying scheme which utilizes ontology schema information for efficient graph search operation. By combining the bidirectional BFS algorithm with run-time class path look up scheme, we could significantly narrow down the search space. The proposed path querying scheme achieved performance improvement by more than 3 orders of magnitude compared with the conventional BFS algorithm. Additionally, because the proposed algorithm works in a scalable manner, it is expected to return graph search results within a reasonable response time on even much larger RDF graph. A drawback we observed in our approach is that there exist empty class path expressions which do not have any instance paths. The empty class path expressions result in unsuccessful search but may cause considerable search cost. For future work, we plan to devise the methods to minimize the empty class path expressions when they are pre-calculated.

References

[1] Huang, J., Abadi, D., Ren, K.: Scalable SPARQL querying of large RDF graphs. Proceedings of the VLDB Endowment (2011)

[2] Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., Stegemann, T.: RelFinder: Revealing relationships in RDF knowledge bases. In: Chua, T.-S., Kompatsiaris, Y., Mérialdo, B., Haas, W., Thallinger, G., Bailer, W. (eds.) SAMT 2009. LNCS, vol. 5887, pp. 182–187. Springer, Heidelberg (2009)

[3] Seo, D., Koo, H.K., Lee, S., Kim, P., Jung, H., Sung, W.-K.: Efficient finding relationship between individuals in a mass ontology database. In: Kim, T.-H., Adeli, H., Ma, J., Fang, W.-C., Kang, B.-H., Park, B., Sandnes, F.E., Lee, K.C. (eds.) UNESST 2011. CCIS, vol. 264, pp. 281–286. Springer, Heidelberg (2011)

[4] Abadi, D., Marcus, A.: Scalable semantic web data management using vertical partitioning. In: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB 2007, pp. 411–422 (2007)

[5] Melton, J., Simon, A.R.: SQL: 1999 - Understanding Relational Language Components

[6] Guo, Y., Pan, Z., Heflin, J.: LUBM: A benchmark for OWL knowledge base systems. Web Semantics Science Services and Agents on the World Wide Web 3(2-3), 158–182 (2005)

Risk Aversion Parameter Estimation for First-Price Auction with Nonparametric Method

Xin An, Jiancheng Chen*, and Yuan Zhang

School of Economics and Management, Beijing Forestry University,
No. 35 Qinghua East Rd., Haidian District, Beijing 100083, P.R. China
{anxin,zhyuan}@bjfu.edu.cn, chenjc1963@163.com

Abstract. More and more clues show that the bidders tend to risk averse. However, traditional nonparametric approach is only applicable for the case of risk neutrality. This study proposes a generalized nonparametric structural estimation procedure for the first-price auctions. To evaluate the performance, extensive Monte Carlo simulation experiments are conducted for ten different values of risk aversion parameter including the risk neutrality case in multiple classic scenes. Though there are no unique estimators for risk aversion parameter, four (weighted) combinations of all estimators are obtained, and some guidance is also given for real-world applications. Finally, empirical results on USFS bidding dataset show that the our nonparametric method can capture bidders' risk aversion to some extend.

Keywords: Risk Aversion, First-Price Auction, Keyword Auction, Monte Carlo, Nonparametric Method.

1 Introduction

Keyword auction [1–3], also known as sponsored search or search auction, is a form of advertising that appears at the top or to the right of the list of results of search engines. Such auctions have become a dominant source of revenue generation on the Internet. To participate in a keyword auction, say for "distributed computing systems", each advertiser submits a cost-per-click (CPC)bid together with a clickable, text-based advertisement. All bids are collected, ranked, and shown automatically in the research result page, and advertisers will pay each time when their advertisements get clicked.

Risk aversion is used to explain the advertisers' behavior under uncertainty. The auction model and the optimal mechanism design for risk averse bidders have been studied by [3–5]. Within the private value paradigm, risk averse bidders tend to shade less their private values relative to the risk neutral case, which often results in some overbidding [6]. More and more clues show that the bidders indeed tend to risk averse [7–9].

In recent years, in order to gain some insight from auction data, structural approaches, pioneered by Paarsch [10], have attracted extensive attention. Some

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing,*
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_39,

of them rely upon a parametric specification of the bidders' private values distribution, e.g., the piecewise pseudo-maximum likelihood estimation (PPMLE) approach [11]. Laffont et al. [12] have proposed a simulated non-linear least square estimation method, which allows for general parametric specifications. Without any parametric assumptions, Guerre et al. [13] have presented a computationally convenient nonparametric estimation procedure. However, the nonparametric approach is only applicable for the case of risk neutrality, which cannot explain the agents' behavior very well. To overcome this problem, this article generalizes nonparametric estimation procedure so that it can be applicable for risk aversion in the first-price auction[14]. Amid the uncertainty of the risk aversion estimator, this study offers several choices and gives some instructive suggestions by contrasting the experimental effects. We then illustrate our choices and suggestions on the US Forest Service timber auctions, and find that the bidders are risk averse.

The rest of the paper is organized as follows. After the risk-averse model for first-price auction with independent private value is briefly introduced in Section 2, a generalized nonparametric estimation procedure is proposed in Section 3. In Section 4 & 5, extensive experimental evaluations on synthetic and real-world data are conducted, and Section 6 concludes this work.

2 The Risk-Averse Model

A single and indivisible object is sold through an auction to N bidders. In this article, the private value paradigm is considered, where every bidder has a private value $v_i (i \in \mathbb{N}_n = \{1, 2, \cdots, n\})$ for the auctioned object. The private values v_is are drawn independently from a distribution $F(\cdot)$, which is assumed to be known to all bidders, and is defined on a compact support $[\underline{v}, \bar{v}]$ with a density $f(\cdot)$. Intuitively, every bidder is potentially risk averse with a von Neuman Morgensten utility function $U(\cdot)$ satisfying $U'(\cdot) > 0, U''(\cdot) \leq 0$ and $U(0) = 0$. The bidders are symmetric in the sense that they share the same private value distribution $F(\cdot)$ and the same utility function $U(\cdot)$. Specifically, the distribution $G(v)$ for the private values is a uniform distribution on $[0, 1]$ with a constant relative risk (CRRA) utility function, $U(x) = x^r, r \in [0, 1]$. In this specification, $1 - r$ is the coefficient of relative risk aversion, with $r = 1$ corresponding to risk neutrality.

Let $b = \sigma(v)$ denote the equilibrium bid function. Under weak regularity conditions, the equilibrium bid function is strictly increasing and differentiable so that its inverse $s(b)$ exists and inherits these properties. Bidder i maximizes his expected utility $\mathbb{E}\Pi_i = [Pr(b_i \geq b_j, j \neq i)]\, U(v_i - b_i) = F(s(b_i))^{N-1}(v_i - b_i)^r$ with respect to his/her bid b_i, where b_j is the j-th player's bid.

The first order condition for maximizing expected utility can be written as

$$(n-1)s'(b_i)f(v_i)(v_i - b_i) - rF(v_i) = 0 \tag{1}$$

To arrange Eq. 1, we obtain

$$v_i = b_i + r \cdot \frac{F(s(b_i))}{f(s(b_i))s'(b_i)(N-1)}. \tag{2}$$

Let $G(b)$ and $g(b)$ be the distribution and density of the bids, respectively. Since $G(b) = F(s(b))$ and $g(b) = f(s(b))s'(b)$, Eq. 2 can be written as

$$v_i = b_i + r \cdot \frac{G(b_i)}{g(b_i)(n-1)}. \tag{3}$$

Additionally, $F(v_i) = v_i$, $f(v_i) = 1$, and $db_i/dv_i = 1/s'(b_i)$, hence from Eq. 1 one can obtain $\frac{db_i}{dv_i} = \frac{n-1}{r}\left(1 - \frac{b_i}{v_i}\right)$. To solve this differential equation, we obtain the equilibrium bid functions

$$b_i \equiv \sigma(v_i) = \frac{n-1}{n-1+r} v_i. \tag{4}$$

3 Generalized Nonparametric Estimation

3.1 Risk Aversion Parameter Estimation

Since one can only access the equilibrium bid and cannot access private value in the real world, the private value v can be seen as a function of equilibrium bid b. In other words, Eq. 1 can also be

$$v_i = b_i + r \cdot \frac{F(\sigma^{-1}(b_i))}{(n-1)f(\sigma^{-1}(b_i))/\sigma'(\sigma^{-1}(b_i))}. \tag{5}$$

From Eq. 4, it is not difficult to see that bid $b = \sigma(v)$ has the same probability distribution function with v, viz. $G(b) = F(\sigma^{-1}(b))$. Guerre et al. [13] have proofed that distribution function $F(\cdot)$ is unique, and equivalents with $G(\cdot)$. Moreover, theirs probability densities also have some relevance, namely $g(b) = f(\sigma^{-1}(b))/\sigma'(\sigma^{-1}(b))$. Therefore, Eq. 5 can be simplified as

$$v_i = b_i + \gamma \cdot \frac{G(b_i)}{g(b_i)(n-1)}. \tag{6}$$

Given T similar single item auctions, which can be divided further into N categories $\mathbb{N}_N$ in terms of the numbers of bidders. Following [9], let v_α and b_α denote the α-th percentile of the distribution $F(\cdot)$ and $G(\cdot)$ respectively, the equation $G(b_\alpha^{(n)}; n) = \alpha = G(b_\alpha^{(l)}; l)$ holds for any different bidders' number $n, l \in \mathbb{N}_N$. Thus, equipped with percentile, Eq. 6 becomes

$$v_\alpha = b_\alpha^{(i)} + \gamma \cdot \frac{G(b_\alpha^{(i)}; i)}{(i-1)g(b_\alpha^{(i)}; i)} = b_\alpha^{(i)} + \frac{\gamma\alpha}{(i-1)g(b_\alpha^{(i)}; i)}, i \in \{n, l\} \tag{7}$$

Through simple arithmetic operations on the above two equations, one can obtain the following estimator $\hat{\gamma}^{(n,l)}$for risk aversion parameter γ,

$$b_\alpha^{(n)} - b_\alpha^{(l)} = \hat{\gamma}^{(n,l)} \cdot \alpha \left(\frac{1}{(l-1)g(b_\alpha^{(l)}; l)} - \frac{1}{(n-1)g(b_\alpha^{(n)}; n)} \right). \tag{8}$$

Of course, $\hat{\gamma}^{(n,l)} = \hat{\gamma}^{(l,n)}(n, l \in \mathbb{N}_N, n \neq l)$. Thus, one can obtain $N(N-1)/2$ estimators for γ. Subsection 3.2 will discuss how to choose a proper one in real-world applications. Surprisingly, $G(\cdot)$ disappears from Eq. 8, which means that the estimators for γ have nothing to do with $G(\cdot)$.

3.2 Choice of Estimators for Risk Aversion Parameters

Intuitively, it seems that some (weighted) combinations of all estimators may be appealing. In this study, two-level combinations are considered. Specially, for any $n \in \mathbb{N}_N (l \in \mathbb{N}_N \backslash n)$, the estimator $\hat{\gamma}^n$ are first obtained from $\hat{\gamma}^{(n,l)}$ using two kinds of weights: $T_l/(T - T_n)$ and $T_l \cdot l/(L_T - T_n \cdot n)$. Here $L_T = \sum_{t=1}^{T} n_t$, $n_t (t \in \mathbb{N}_T)$ is the number of bidders in the t-th auctions, T_l and T_n are the times of auctions with the number of bidders is l and n, respectively. Thus, there are two inner estimators for γ^n. Similarly, the outer combination also using two kinds of weights: T_n/T and $T_n \cdot n/L_T$. So, there are four estimators for the risk aversion parameter γ in total. See Table 1 for details.

Table 1. Four Estimators for γ

$\hat{\gamma}^{a,a}$: $\sum_{n\in\mathbb{N}} \frac{T_n}{T} \sum_{l\in\mathbb{N}_N\backslash n} \frac{T_l}{T-T_n} \hat{\gamma}^{(n,l)}$	$\hat{\gamma}^{a,b}$: $\sum_{n\in\mathbb{N}} \frac{T_n}{T} \sum_{l\in\mathbb{N}_N\backslash n} \frac{T_l \cdot l}{L_T - T_n \cdot n} \hat{\gamma}^{(n,l)}$
$\hat{\gamma}^{b,a}$ $\sum_{n\in\mathbb{N}} \frac{T_n \cdot n}{L_T} \sum_{l\in\mathbb{N}_N\backslash n} \frac{T_l}{T-T_n} \hat{\gamma}^{(n,l)}$	$\hat{\gamma}^{b,b}$: $\sum_{n\in\mathbb{N}} \frac{T_n \cdot n}{L_T} \sum_{l\in\mathbb{N}_N\backslash n} \frac{T_l \cdot l}{L_T - T_n \cdot n} \hat{\gamma}^{(n,l)}$

There is no apparent reasons to prefer one estimator to another, and many factors may influence one's choice; such as, auctions' number, the number of bidders, etc. In this situation, Monte Carlo experiments (see Section 4) can be utilized to examine the performance of different estimators. Before this, given observable bids $\{b_{i,t}, i \in \mathbb{N}_n; t \in \mathbb{N}_T\}$, a simple structural estimation procedure is summarized as follows:

STEP 1: To estimate the distribution function $G(\cdot)$ and density function $g(\cdot)$ of observed bids as follows: $\hat{G}(b; n) = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{1}(b_{i,t} \leq b)$, and $\hat{g}(b; n) = \frac{1}{nTh_g} \sum_{t=1}^{T} \sum_{i=1}^{n} K_g \left(\frac{b - b_{i,t}}{h_g} \right)$, where $\mathbf{1}(\cdot)$ is a indicator function, $K_g(\cdot)$ is a kernel with a compact support and h_g is a vanishing bandwidth.

STEP 2: To calculate four estimators for the risk aversion parameter according to Table 1.

STEP 3: To estimate the private valuation $\{\hat{v}_{i,t}, i \in \mathbb{N}_n; t \in \mathbb{N}_T\}$ by Eq. 6.

STEP 4: To estimate the density function of private valuation $f(v; n)$ as follows: $\hat{f}(v; n) = \frac{1}{nTh_f} \sum_{t=1}^{T} \sum_{i=1}^{n} K_f \left(\frac{b - b_{i,t}}{h_f} \right)$, where h_f is a vanishing bandwidth, $K_f(\cdot)$ is a kernel function.

4 Experimental Design and Discussions

4.1 Experimental Design

In order to evaluate the performance of all estimators, extensive Monte Carlo experiments are conducted in this section. In all the simulation experiments, the

private costs are draw from a uniform distribution on $[0, 1]$, and the bidders have a CRRA utility function, $U(x) = x^r, r \in [0, 1]$. The experiment takes ten different parameter values of relative risk aversion: $\gamma_i = 0.1 \times i (i = 1, \cdots, 10)$ with $\gamma_{10} = 1.0$ corresponding to risk neutrality. In order to simulate the real auction data as closely as possible, the experiment considers three kind of sample sizes T: 400, 2000 and 8000, which corresponds to small, median and large sample, respectively. In each of these samples, the number of bidder n could take on four different values: 3, 6, 9 and 12. We investigate three different patterns for the design matrix and the probability distribution of the ns. Table 2 illustrates the detailed T_ns and their corresponding $p(n)$s for the three different designs.

Table 2. Design Matrices of the T_ns

Sample Size		400				2000				8000			
n		3	6	9	12	3	6	9	12	3	6	9	12
Design A	T_n	100	100	100	100	500	500	500	500	2000	2000	2000	2000
	$p(n)$	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Design B	T_n	40	80	120	160	200	400	600	800	800	1600	1800	3200
	$p(n)$	0.10	0.20	0.30	0.40	0.10	0.20	0.30	0.40	0.10	0.20	0.30	0.40
Design C	T_n	160	120	80	40	800	600	400	200	3200	2400	1600	800
	$p(n)$	0.40	0.30	0.20	0.10	0.40	0.30	0.20	0.10	0.40	0.30	0.20	0.10

To obtain the simulated auction data, we first generate randomly private valuations for different designs from the uniform distribution. We then compute numerically bids using Eq. 4 with different values for γ. Note that the random numbers for the experiments are generated using the multiplicative congruential method with modulus $(2^{31}-1)$, multiplier 397,204,094, and initial seed 2,420,375. To minimize the impact of noise, the preceding procedure is repeated 1000 times.

In our experiments, let $K(u) = (35/32)(1-u^2)^3\mathbf{1}(|u| \leq l), h_g = 1.06\hat{\sigma}_b(nT)^{-0.2}$, $h_f = 1.06\hat{\sigma}_v(nT)^{-0.2}$, where $\hat{\sigma}_b$ and $\hat{\sigma}_v$ are the standard deviations (STD) of observed bid and private valuation, respectively. However, the kernel density estimator is biased at the boundaries of the support. Similar to [13], 10% observed pseudo private values near the boundaries are trimmed.

4.2 Results and Discussions

From Eq. 8, it is not difficult to see that the percentile α may influences the estimation for γ. In order to minimize the influence α on risk aversion estimator, the percentile α is optimized as follows: select α^* from $\{0.9, 0.91, \cdots, 0.99\}$, so that STD of estimators for γ is minimized. Table 3 reports the estimated values for γ with α^*. One can see that that all the means of risk aversion estimator are near to the real values, especially for large sample. This means that our simple structure estimation procedure is feasible.

In order to improve these estimators' usability, Table 4 gives some useful suggestion based on $\arg\min_{x,y\in\{a,b\}} \sum_i |\hat{\gamma}^{x,y} - 0.1 \times i|/(0.1 \times i)$. To assess the goodness of fit of the structural estimation procedure, we also calculate the 2-norm between the estimated and actual private values, defined formally as $d = \sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{v}_{i,t} - v_{i,t})^2}$, where $\hat{v}_{i,t}$ and $v_{i,t}$ denote the bidder i's private

Table 3. Estimated Values for 4 Estimators for γ in Design A, B & C

Sample Size T	Estimator	Mean (Design A)			Mean (Design B)			Mean (Design C)		
		γ_2	γ_5	γ_{10}	γ_2	γ_5	γ_{10}	γ_2	γ_5	γ_{10}
400	$\hat{\gamma}^{a,a}$	0.1837	0.5008	1.0019	0.1835	0.5003	1.0105	0.1860	0.4859	0.9946
	$\hat{\gamma}^{a,b}$	0.1807	0.5037	1.0011	0.1935	0.4958	1.0243	0.1914	0.4868	0.9996
	$\hat{\gamma}^{b,a}$	0.1816	0.5024	1.0004	0.1892	0.4971	1.0176	0.1900	0.4865	0.9979
	$\hat{\gamma}^{b,b}$	0.1778	0.5074	1.0030	0.2010	0.4929	1.0354	0.1995	0.4886	1.0045
2000	$\hat{\gamma}^{a,a}$	0.1863	0.4701	0.9468	0.1876	0.4745	0.9570	0.1840	0.4663	0.9392
	$\hat{\gamma}^{a,b}$	0.1861	0.4704	0.9481	0.1871	0.4745	0.9577	0.1830	0.4670	0.9443
	$\hat{\gamma}^{b,a}$	0.1861	0.4702	0.9474	0.1873	0.4743	0.9569	0.1833	0.4668	0.9429
	$\hat{\gamma}^{b,b}$	0.1860	0.4711	0.9503	0.1867	0.4746	0.9587	0.1819	0.4682	0.9517
8000	$\hat{\gamma}^{a,a}$	0.1842	0.4620	0.9262	0.1837	0.4622	0.9276	0.1836	0.4607	0.9235
	$\hat{\gamma}^{a,b}$	0.1839	0.4619	0.9264	0.1833	0.4620	0.9276	0.1837	0.4612	0.9248
	$\hat{\gamma}^{b,a}$	0.1840	0.4619	0.9262	0.1835	0.4620	0.9274	0.1837	0.4611	0.9245
	$\hat{\gamma}^{b,b}$	0.1836	0.4619	0.9268	0.1830	0.4618	0.9278	0.1840	0.4621	0.9268

value and actual value in auction t, respectively. These results are reported in Table 5 where the estimator for γ is set by Table 4. From Table 5, one can easily see that private values can be estimated very well, especially for large sample.

Table 4. Design Matrices of the T_ns

Sample Size	Design A	Design B	Design C
small	$\hat{\gamma}^{a,b}$	$\hat{\gamma}^{a,a}$	$\hat{\gamma}^{a,a}$
medium	$\hat{\gamma}^{a,a}$	$\hat{\gamma}^{b,b}$	$\hat{\gamma}^{a,b}$
large	$\hat{\gamma}^{a,b}$	$\hat{\gamma}^{a,b}$	$\hat{\gamma}^{a,a}$

Table 5. The 2-Norm between the Estimated and Actual Private Values

Sample Size		Design A			Design B			Design C		
		400	2000	8000	400	2000	800	400	2000	8000
Mean	γ_2	0.0163	0.0084	0.0036	0.0103	0.0059	0.0031	0.0194	0.0084	0.0049
	γ_5	0.0224	0.0118	0.0073	0.0235	0.0089	0.0060	0.0238	0.0132	0.0093
	γ_{10}	0.0391	0.0188	0.0128	0.0308	0.0151	0.0106	0.0432	0.0093	0.0158

5 Empirical Application

USFS (US Forest Service) biding dataset covers the western half, divided into 9 regions, of the US, which contains 13,318 first-price sealed-bid auctions during 1974 to 1993. The data provides detailed information on the estimated volume of timber, the number of acres of the parcel, the estimated appraisal value of the timber, the exact location of the timber parcel, the term of the contract, the logging costs as well as other costs such as road construction costs, the number of bidders who have submitted a bid as well as their bids in dollars and their identity. Table 6 gives some summary information on the total bids. Since it is well accepted among economists that the reserve price in USFS does not act as a screening device as it is set too low, the reserve price is nonbinding.

Table 6. The Distance between the Estimated and Actual Private Values

	Bids ($)	Winding Bids ($)	Appraisal Value ($ per mbf)	Volume (mbf)	Number of Bidders
Mean	8599.12	9465.33	13297.80	2312.16	3.93
STD	7398.31	7923.36	12664.30	3549.23	2.10

By Subsection 4.2, the optimal percentile $\alpha^* = 0.94$. Additionally, since USFS dataset is a large sample and large ns were more likely than small ones, it is better to adopt the estimator $\hat{\gamma}^{a,a}$ in this case by Table 4. Following simple structural estimation procedure in Subsection 3.2, one can obtain $\hat{\gamma}^{a,a} = 0.5866$, which falls in between 0.3187 in [9] and 0.6813 in [8]. It is very possible that [8] overestimated γ and [9] underestimated γ.

6 Conclusions

A generalized nonparametric structural estimation procedure is proposed for risk-aversion case in the first-price auctions. In order to evaluate the performance, extensive Monte Carlo simulation experiments are conducted for ten different values of γ including the risk-neutral case in multiple classic scenes. Though there are no unique estimators for risk aversion parameter, four (weighted) combinations of all estimators are obtained, and some guidance is also given for real-world applications. Finally, empirical results on bidding data from the timber sales at the USFS show that the our nonparametric method can capture bidders' risk aversion to some extend. What's more, our method can be used easily in a secure trusted auction oriented clustering, such as STACRP [14].

Acknowledgments. This work was funded by "Fundamental Research Funds for the Central Universities" (TD2012-08) and Beijing Forestry University Young Scientist Fund(BLX2011028).

References

1. Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: Estimating the click-through rate for new ads. In: WWW 2007, pp. 521–529. ACM, New York (2007)
2. Varian, H.R.: Position auctions. Int. J. Ind. Organ. 25(6), 1163–1178 (2007)
3. Goel, A., Munagala, K.: Hybrid keyword search auctions. In: WWW 2009, pp. 221–230. ACM, New York (2009)
4. Maskin, E., Riley, J.: Optimal auctions with risk averse buyers. Econometrica 52(6), 1473–1518 (1984)
5. Matthews, S.: Comparing auctions for risk averse buyers: A buyer's point of view. Econometrica 55(3), 633–646 (1987)
6. Cox, J.C., Smith, V.L., Walker, J.M.: Theory and individual behavior of first-price auctions. J. Risk Uncertain. 1(1), 61–99 (1988)
7. Goeree, J.K., Holt, C.A., Palfrey, T.R.: Quantal response equilibrium and overbidding in private value auctions. J. Econ. Theory 104(1), 247–272 (2002)
8. Lu, J., Perrigne, I.: Estimating risk aversion from ascending and sealed-bid auctions: the case of timber auction data. J. Appl. Econometr. 23(7), 871–896 (2008)
9. Campo, S., Guerre, E., Perrigne, I., Vuong, Q.: Semiparametric estimation of first-price auctions with risk-averse bidders. Rev. Econ. Stud. 78(1), 112–147 (2011)
10. Paarsch, H.J.: Deciding between the common and private value paradigms in empirical models of auctions. J. Econometr. 51(1-2), 191–215 (1992)

11. Donald, S., Paarsch, H.J.: Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. Econometric Theory 12(3), 517–567 (1996)
12. Laffont, J.J., Ossard, H., Vuong, Q.: Econometrics of first-price auctions. Econometrica 34(4), 953–980 (1995)
13. Guerre, E., Perrigne, I., Vuong, Q.: Optimal nonparametric estimation of first-price auctions. Econometrica 68(3), 525–574 (2000)
14. Chatterjee, P., Sengupta, I., Ghosh, S.: STACRP: A secure trusted auction oriented clustering based routing protocol for MANET. Cluster Comput. 15(3), 303–320 (2012)

Diverse Heterogeneous Information Source-Based Researcher Evaluation Model for Research Performance Measurement

Jinhyung Kim[1], Myunggwon Hwang[1], Do-Heon Jeong[1], Sa-kwang Song[1], Jangwon Gim[1], Hanmin Jung[1], Shuo Xu[2], and Lijun Zhu[2]

[1] Dept. of Computer Intelligence Research,
Korea Institute of Science and Technology Information,
52-11, Eoeun dong, Yuseong gu, Daejeon, Republic of Korea
[2] Information Technology Supporting Center,
Institute of Scientific and Technical Information of China,
No. 15 fuxing Rd., Haidian District, Beijing 100038, P.R. China
{jinhyung,mgh,heon,esmallj,jangwon,jhm}@kisti.re.kr,
{xush,zhulj}@istic.ac.cn

Abstract. Analysis, prediction, and recommendation of information about experts are very important tasks for future research planning and strategy establishment. However, it takes much time and efforts even for precise analysis of experts because we need to analyze huge and diverse heterogeneous information. There are several application and tools for supporting analysis about researchers, but they provides fragmentary analysis result based on simple evaluation criteria. Therefore, in this paper, we suggest new researcher evaluation model based on diverse performance evaluation features, named RSW model. By using RSW model, we can analyze and compare researchers in various perspectives. In addition, we can ranked researchers and recommend outstanding collaborator in a specified research field.

Keywords: Researcher Performance Evaluation, Evaluation Features, Researcher Evaluation Model, Performance Measurement.

1 Introduction

Analyzing and discovering experts is very critical part in research planning, research strategy establishing, and collaborator/competitor finding. Therefore, several counseling company and government institute have been researching for discovering experts by analyzing diverse related information such as arnetminer, MS academic search, Google scholar, and so on [1-6]. However, most of existing applications or tools analyze researchers in fragmentary criteria such as number of papers or patents or citation. Analysis results by the existing applications can not be guaranteed by all of researchers because researchers can have various analysis purpose. For example, strategy establisher in some company can be interested in commercial aspect of experts but professor in some university can have interest in scholar aspect of

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_40,

researchers. Therefore, in this paper, we newly define seven features for analyzing researchers in diverse aspects. By using diverse features, we can analyze research capability and performance of researchers and compare researchers in several aspects. In addition, we can suggest researcher ranking based on analysis results of diverse features.

2 RSW Model

In this section, we describe researcher evaluation model named RSW model based on several measurement features such as commerciality, scholarity, and so on. Each measure feature can be presented as combination of data element defined in metadata from literature resources such as papers, patents, web reports, and SNS data. Measure features of RSW model consists of 7 features; Commerciality, Scholarity, Sociality, Interest, Influentiality, Diversity, Durability. RSW model is focusing on analyzing internal capability of researchers; how many papers were written by researchers, how many citations were acquired, and so on.

Table 1. Description of acronym

Acronym	*Description*
$CIT_t^K(PR_n)$	Average number of patent citation of person PR_n related K topic at t year
$PTA_t^K(PR_n)$	Number of applied patent of person PR_n related K topic at t year
$PTR_t^K(PR_n)$	Number of registered patent of person PR_n related K topic at t year
$CPP_t^K(PR_n)$	Patent CPP index of person PR_n related K topic at t year
$HID_t^K(PR_n)$	H-index of person PR_n related K topic at t year
$GID_t^K(PR_n)$	G-index of person PR_n related K topic at t year
$PPC_t^K(PR_n)$	Number of proceeding papers of person PR_n related K topic at t year
$PPJ_t^K(PR_n)$	Number of Journal papers of person PR_n related K topic at t year
$PAR_t^K(PR_n)$	Number of positive articles in social data of person PR_n related K topic at t year
$NAR_t^K(PR_n)$	Number of negative articles in social data of person PR_n related K topic at t year
$EDC_t^K(PR_n)$	Number of conference editors Number of positive articles in social data of person PR_n related K topic at t year
$EDJ_t^K(PR_n)$	Number of journal editors Number of positive articles in social data of person PR_n related K topic at t year
$WN_t^K(PR_n)$	Number of web articles of person PR_n related K topic at t year
$WR_t^K(PR_n)$	Number of web reports of person PR_n related K topic at t year

Table 1. (*continued*)

$PC_t^K(PR_n)$	Number of paper citation of person PR_n related K topic at t year
$PCoA_t^K(PR_n)$	Number of co-authors of person PR_n related K topic at t year
$PCA_t^K(PR_n)$	Number of correspondence authors of person PR_n related K topic at t year
$PTCoA_t^K(PR_n)$	Number of co-developer of person PR_n related K topic at t year
$PCaT_t^K(PR_n)$	Number of paper categories of person PR_n related K topic at t year
$PK_t^K(PR_n)$	Number of paper keywords of person PR_n related K topic at t year
$PTIPC_t^K(PR_n)$	Number of IPC class of person PR_n related K topic at t year
$WCaT_t^K(PR_n)$	Number of web article categories of person PR_n related K topic at t year
$WK_t^K(PR_n)$	Number of keywords extracted from SNS articles of person PR_n related K topic at t year

(Definition 1: Commerciality) Commerciality can be defined as ability to produce practical products and profits. Therefore, commerciality is closely related to patent in literature resources. Commerciality is combination of number of applied patents, number of registered patents, and CPP-index.

$$Commercial\,ity_t^K(PR_1) = \frac{\frac{CIT_t^K(PR_1)}{PTR_t^K(PR_1)} * \{PTA_t^K(PR_1) + PTR_t^K(PR_1)\}}{\sum_{l=1}^{n}\{PTA_t^{K_l}(PR_1) + PTR_t^{K_l}(PR_1)\}}$$

$$= \frac{\frac{CIT_t^K(PR_1) * PTA_t^K(PR_1)}{PTR_t^K(PR_1)} + \frac{CIT_t^K(PR_1) * PTR_t^K(PR_1)}{PTR_t^K(PR_1)}}{\sum_{l=1}^{n}\{PTA_t^{K_l}(PR_1) + PTR_t^{K_l}(PR_1)\}}$$

$$= \frac{\frac{CIT_t^K(PR_1) * PTA_t^K(PR_1)}{PTR_t^K(PR_1)} + CIT_t^K(PR_1)}{\sum_{l=1}^{n}\{PTA_t^{K_l}(PR_1) + PTR_t^{K_l}(PR_1)\}}$$

$$= \frac{CIT_t^K(PR_1) * \{1 + \frac{PTA_t^K(PR_1)}{PTR_t^K(PR_1)}\}}{\sum_{l=1}^{n}\{PTA_t^{K_l}(PR_1) + PTR_t^{K_l}(PR_1)\}}$$

(Definition 2: Scholarity) Scholarity can be defined as ability to produce new knowledge and academic output. Therefore, scholarity is closely related to paper in literature resources. Scholarity is combination of number of conference/journal papers, h-index, and g-index.

$$Scholarity_t^K(PR_1) = \frac{HID_t^K(PR_1) * \{PPC_t^K(PR_1) + PPJ_t^K(PR_1)\}}{\sum_{l=1}^{n}\{PPC_t^{K_l}(PR_1) + PPJ_t^{K_l}(PR_1)\}}$$
$$= \frac{HID_t^K(PR_1) * PPC_t^K(PR_1) + HID_t^K(PR_1) * PPJ_t^K(PR_1)}{\sum_{l=1}^{n}\{PPC_t^{K_l}(PR_1) + PPJ_t^{K_l}(PR_1)\}}$$

(Definition 3: Sociality) Sociality can be defined as ability to create social relationship and tendency to form communities and societies. Sociality is combination of number of SNS articles, positive-negative index of articles, and number of journal/conference editors.

$$Sociality^K(PR_1)$$
$$= \frac{\sum_{t=2001}^{2012}[\{(+1) * PAR_t^K(PR_1) + (-1) * NAR_t^K(PR_1)\} + \{EDC_t^K(PR_1) + EDJ_t^K(PR_1)\}]}{\sum_{l=1}^{n}[\sum_{t=2001}^{2012}\{AR_t^K(PR_l) + EDC_t^K(PR_l) + EDJ_t^K(PR_l)\}]}$$

(Definition 4: Interest) Interest can be defined as attraction and concern by other researchers. Interest is combination of number of web articles/reports, SNS articles, and positive-negative index of articles.

$$Interest_t^K(PR_1) =$$
$$\frac{\sum_{t=2001}^{2013}[\{(+1) * PAR_t^K(PR_1) + (-1) * NAR_t^K(PR_1)\} + \{WN_t^K(PR_1) + WR_t^K(PR_1)\}}{\sum_{l=1}^{n}[\sum_{t=2001}^{2012}\{AR_t^K(PR_l) + WN_t^K(PR_1) + WR_t^K(PR_1)\}}$$

(Definition 5: Influentiality) Influentiality can be defined as ability to spread leverage to other researchers. Influentiality is combination of number of paper citation, co-authors, correspondence authors, patent CPP-index, and number of co-developer in patent.

$$Influentiality^{K}(PR_1) = \frac{\sum_{t=2001}^{2013}\{PC_t^K(PR_1)+PCoA_t^K(PR_1)+PCA_t^K(PR_1)+CPP_t^K(PR_1)+PTCoA_t^K(PR_1)}{\sum_{l=1}^{n}\{\sum_{t=2001}^{2013}PC_t^K(PR_l)+CPP_t^K(PR_l)+3l\}}$$

(Definition 6: Diversity) Diversity can be defined as ability to extend research scope and evaluated with degree of variation of research field. Diversity is combination of number of paper categories, paper keywords, patent IPC class, categories in web resources, and number of keyword extracted from SNS articles.

$$Diversity^{K}(PR_1) = \frac{\sum_{t=2001}^{2013}\{PCaT_t^K(PR_1)+PK_t^K(PR_1)+PTIPC_t^K(PR_1)+WCaT_t^K(PR_1)+SK_t^K(PR_1)\}}{\sum_{l=1}^{n}[\sum_{t=2001}^{2013}\{PCaT_t^K(PR_l)+PK_t^K(PR_l)+PTIPC_t^K(PR_l)+WCaT_t^K(PR_l)+SK_t^K(PR_l)\}]}$$

(Definition 7: Durability) Durability can be defined as ability to keep research consistently in some specified research field. Durability is combination of number of conference/journal papers, applied/registered patents, web articles, and web reports.

$$if\{(PTA_t^K(PR_1)+PTR_t^K(PR_1)==0\}$$
$$DPT_t^K = 0 / else / DPT_t^K = 1$$
$$if\{(PPC_t^K(PR_1)+PPJ_t^K(PR_1)==0\}$$
$$DPT_t^K = 0 / else / DPT_t^K = 1$$
$$if\{(WN_t^K(PR_1)+WR_t^K(PR_1)==0\}$$
$$DW_t^K = 0 / else / DW_t^K = 1$$

$$Durability_t^K(PR_1) = \frac{\sum_{t=2001}^{2013}(DPT_t^K + DPP_t^K + DW_t^K)}{39}$$

3 Conclusion

In this paper, we suggest the evaluation model about performance of researchers based on various evaluation features, named RSW model. In addition, we define 7

evaluation features by combination of data elements extracted by literature information such as papers, patents, web resources, and SNS articles: (1) commerciality, (2) scholarity, (3) sociality, (4) interest, (5) Influentiality, (6) diversity, (7) durability. By the RSW model, we can analyze and compare researchers in diverse perspective, not just fragmentary analysis with simple single criteria.

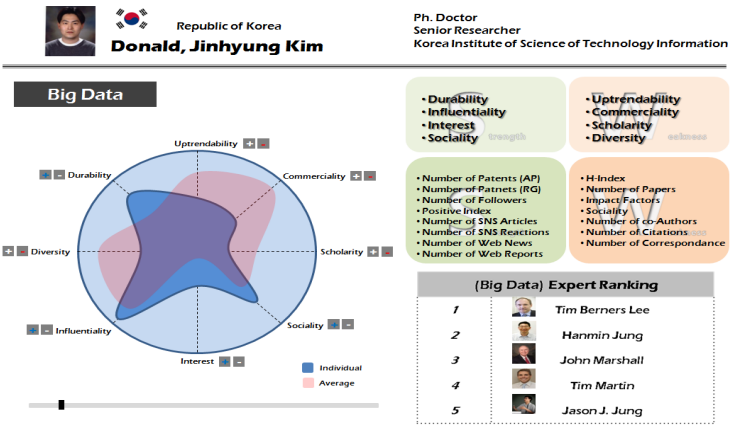


Fig. 1. Snapshot of Researcher Analysis Service based on RSW Model

References

1. Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., Li, J.: ArnetMiner: An Expertise Oriented Search System for Web Community. Frontiers of Computer Science in China 2(1), 94–105 (2008)
2. Hamaker, C., Apry, B.: Google Scholar. Serials 18(1), 70–72 (2005)
3. Noruzi, A.: Google Scholar: The New Generation of Citation Indexes. International Journal of Libraries and Information Services 55(4), 170–180 (2005)
4. Philipp, M., Kathrin, W.: An Exploratory Study of Google Scholar. Online Information Review 31(6), 814–830 (2007)
5. Hands, A.: Microsoft Academic Search. Technical Services Quarterly 29(3), 251–252 (2012), `http://academic.research.microsoft.com`
6. Peter, J.: The Pros and Cons of Microsoft Academic Search from a Bibimetric Perspective. Online Information Review 35(6), 983–997 (2011)

Fast Big Textual Data Parsing in Distributed and Parallel Computing Environment

Jung-Ho Um, Chang-Hoo Jeong, Sung-Pil Choi, Seungwoo Lee, and Hanmin Jung

Dept. of Computer Intelligence Research,
Korea Institute of Science and Technology Information,
245 Daehakno, Yuseong-gu, Daejeon, 305-806, Korea
{jhum,chjeong,spchoi,swlee,jhm}@kisti.re.kr

Abstract. Currently, tremendous numbers of scientific and technical articles are being published due to the rapid development of the scientific and technical fields. Also, systems are being proposed which can give useful information to users by extracting information from scientific and technical articles. For such systems, we need to be able to extract information from a massive number of documents very fast and reliably. However, legacy parsers, such as Stanford, Enju and so on, cannot consider a large number of documents because such parsers analyze wide context range of the sentence for their parsing, and so those parsers require a lot of time to run. Therefore, in this paper, we report on the development of a parser which is based on MapReduce, a distributed and parallel programming model. Our parser has achieved about nineteen times better performance than that of one of the-state-of-the-art legacy parsers.

Keywords: distributed and parallel computing, big textual data, parsing, MapReduce.

1 Introduction

Currently, tremendous numbers of scientific and technical articles are being published due to the rapid development of the scientific and technical fields. Also, systems are being proposed which can give useful information to users by extracting information from scientific and technical articles [1]. To extract meaningful information from documents or articles, such systems require natural language processing. Parsing is an essential part of natural language processing; it is used for recognized relations between subjects and verbs or objects and verbs.

However, parsing requires a lot of time to run for a large number of documents because it consider semantics for wide context range of the sentence to enhance precision and recall value. This type of processing proportionally affects the execution time with increasing in the number of documents. Therefore, a parsing system running on a distributed and parallel environment needs to be developed to parse massive numbers of documents concurrently. In this paper, we propose a parsing system that applies a Stanford parser to the MapReduce framework order to extract information

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_41,

very fast. The reason why we use the Stanford parser is that it is one of the-state-of-the-art parsers and has high precision and recall in parsing documents.

The paper is organized as follows. Section 2 introduces related work. We report on the design of the proposed distributed and parallel text parsing system in Section 3. Section 4 describes the experimental results. Finally, in Section 5, we give the conclusion.

2 Related Works

In this section, we introduce the Stanford parser [2] and the MapReduce framework [3]. These are used by the proposed system in distributed and parallel environments. First, the Stanford parser, proposed by the NLP lab of Stanford University in the 1990s, enhances precision and optimizes the performance by using the PCFG model [2]. The Stanford parser is currently released as an open source program. The Stanford parser analyzes sentences using the PCFG model and notes the subjects, objects, and related verbs as a form of dependency tree structure. However, many studies that have considered the Stanford parser focus on enhancing the performance of the parsing algorithm.

On the other hand, the MapReduce framework is a parallel programming model proposed by Dean and Ghemawat in 2003[3]. It consists of a user-defined Map function and a Reduce function. These two functions reside on each server, in order to allow the servers to process data in parallel (See Figure 1).

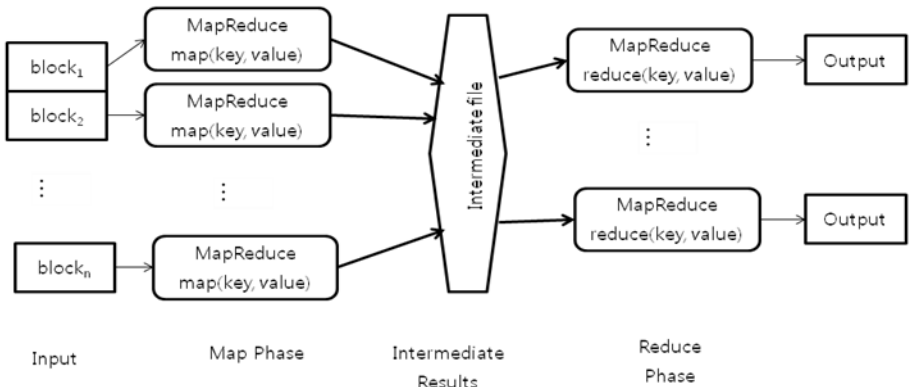


Fig. 1. The MapReduce framework

As can be seen in Figure 1, the input data is equally split ant then assigned to each server. The Map function processes data locally on each of the servers and the Reduce function merges the computed data following a user-defined process.

3 Proposed Distributed and Parallel Parsing System

In order to extract information from massive numbers of documents, we have designed a distributed and parallel textual parser which is implemented by applying the Stanford parser to the MapReduce framework. The proposed system architecture is shown in Figure 2.

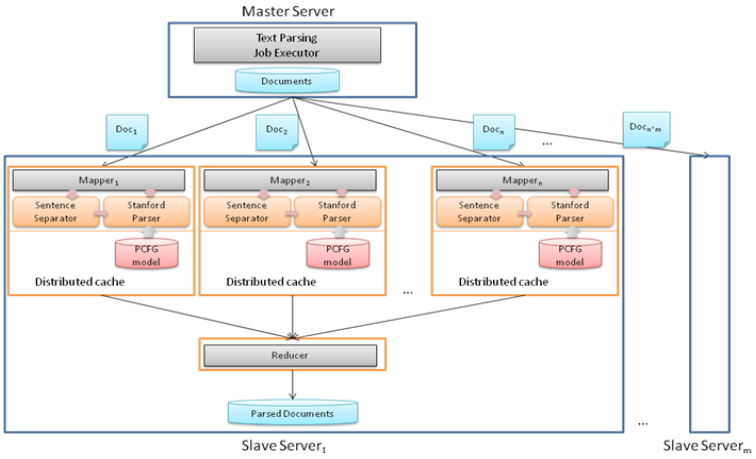


Fig. 2. The system architecture

The system stores the input and output data, such as documents and parsed sentences to the hadoop file system [4]. The PCFG model that is used by the Stanford parser is also loaded into the hadoop file system because the PCFG model is necessary for all mappers, allowing them to parse sentences. Therefore, the PCFG model is loaded as a distributed cache for sharing all of the mappers.

The work flows of the mapper and the reducer are described as follows. First, the mapper separates documents into sentences and then parses the sentences by calling the PCFG parsing module of the Stanford parser. At that time, the Stanford parser needs the PCFG model. For this reason, the PCFG model is stored in the distributed cache. After the completion of parsing, the map function writes the parsed sentence information in a dependency tree form. This format can be used to recognize relations between words or phrases. Next the reducer merges the dependency tree structure for each sentence and then writes final results. Algorithm 1 shows the pseudo codes of the Mapper and Reducer class.

Proposed system has three advantages. First, it reduces the parsing time because the system analyzes massive number of documents concurrently on distributed and parallel environments, while legacy parsers require a lot of parsing time because they analyze documents sequentially. Second, the proposed system can maintain the same high precision performance as the Stanford parser applied to the proposed system. Finally, the system has high portability. The reason for this is that if users want to change the parser, the system can easily be changed by modifying only the parser calling API part with replacing the Stanford parser with the legacy parser.

```
1:  Class Mapper
2:    Method Setup()
3:      DistributedCache.add(PCFGmodel)
4:      StanfordParser.setModel(PCFGmodel)
5:    Method Map(document d)
6:      sentence <- SetenceSeparator(d)
7:      dependencyTree <- Parser.parsePCFG(sentence)
8:      emit(d.id, dependencyTree)


1: Class Reducer
2:    Method Reduce(d.id, Iterable<dependencyTree>)
3:        for each dependecyTree dt
4:             output+=dt
5:         emit(d.id, output)
```

Algorithm 1. Pseudo code

4 Experimental Results

To evaluate the performance of the proposed system, we consider two different HW environments such as single server and nineteen servers. On the single server, we compare the running time of the legacy Stanford parser with that of the distributed Stanford parser applied to the MapReduce framework. We also evaluate the running time of the distributed Stanford parser on nineteen servers. Servers consist of eight cores of Intel i7, 32GByte memories and 2TB storages; we use hadoop 0.20.203 and the Stanford Parser 2.0.4. The data set consists of 10,000 paper abstracts from NDSL owned by KISTI. We evaluate the running time from the step of the sentence separation to the step of parsing sentences using the distributed Stanford parser. The experimental results are shown in Figure 3. The java heap memory size for the experiments is set to 4 GB. For the single server, the execution time of the proposed system is almost half that of the legacy Stanford parser, when the numbers of mappers and reducers are four and one, respectively. Even though the system uses four mappers, the performance gain is only double. The reason for this is that the system requires starting time to initialize the MapReduce framework and additional time to merge the results. In this experiment, we found that the speed is proportional to half the number

of mappers. The results for the distributed Stanford parser, where that parser is running on 19 servers, follow our predictions. This means that the proposed system has an advantage in terms of scalability.

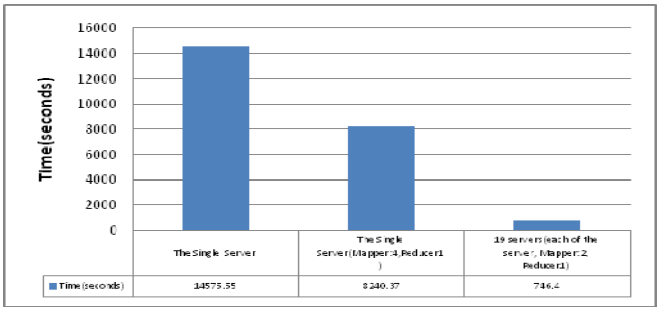


Fig. 3. Experimental results

5 Conclusion

In this paper, we propose a big textual parsing system which applies the Stanford parser to the MapReduce framework in distributed and parallel environments. As a result, the experimental results show that the proposed system has the advantage of scalability. Proposed parser contributes that it can parse massive number of documents by adapting Stanford parser to distribute and parallel environments. Meanwhile, legacy parsers consume a lot of time to parse because they parse sentence by sentence sequentially.

For future work, we will evaluate the system by using two million documents collected from the works in 2012 [5, 6]. In addition, we will study the optimization of the parser in distributed and parallel environments.

References

1. Kim, J., Lee, S., Jeong, D.-H., Jung, H.: Semantic Data Model and Service for Supporting Intelligent Legislation Establishment. In: The 2nd Joint International Semantic Technology Conference (2012)
2. Klein, D., Manning, C.D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423–430 (2003)
3. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: OSDI, pp. 137–150 (2004)
4. HDFS (hadoop distributed file system) architecture (2009), `http://hadoop.apache.org/common/docs/current/hdfs-design.html`
5. Shin, S., Um, J., Song, S.-K., Choi, S.-P., Jung, H.: uLAMP: unified Linguistic Assets Management System. In: The 2nd Joint International Semantic Technology Conference (2012)
6. Seo, D., Hwang, M.-N., Shin, S., Choi, S.: Development of Crawler System Gathering Web Document on Science and Technology. In: The 2nd Joint International Semantic Technology Conference (2012)

K-Base: Platform to Build the Knowledge Base for an Intelligent Service

Sungho Shin[1], Jung-Ho Um[1], Sung-Pil Choi[1], Hanmin Jung[1], Shuo Xu[2], and Lijun Zhu[2]

[1] Department of Computer Intelligence Research, Korea Institute of Science and Technology Information, Korea, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea
[2] Information Technology Supporting Center, Institute of Scientific and Technical Information of China, No. 15 fuxing Rd., Haidian District, Beijing 100038, P.R. China
{maximus74,jhum,spchoi,jhm}@kisti.re.kr, {xush,zhulj}@istic.ac.cn

Abstract. Recently, there is an increasing interest in effectively using big data. It is also thought that the machine learning methods are crucial to effectively extract knowledge from big text data when they are coupled with big data technologies such as MapReduce and Hadoop. For tasks such as the knowledge extraction from huge amount of texts and the reasoning, it produces better results to simultaneously apply a machine learning method and big data technologies to the system. In this research, we propose a system using a machine learning method and big data technologies, and compare it with the existing system in terms of velocity and accuracy. The proposed system is expected to faster and more accurately build the knowledge base than the existing system.

Keywords: Distributed and parallel computing, Knowledge base, Machine learning, Knowledge extraction, Reasoning.

1 Introduction

The knowledge extraction is a type of information extraction. If the information extracted from texts is well organized with each other, it evolves to and gives the knowledge used for supporting human's decision. What is significantly considered in designing a system building knowledge base is the optimization of the process and the method applied to achieve goals of the system.

In this research, we analyze an existing system building knowledge base and find some improvements for the goal of better performance in terms of velocity of system and the accuracy of output data. The existing system was implemented to run on single machine and use dictionaries and rules to building knowledge base so that it is not appropriate for processing huge amount of texts like big data in terms of velocity and accuracy. In order to improve the existing system, we suggest a new knowledge

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_42,

extraction system having a better process to build knowledge base and architecture based on the process. The main idea of the new process and architecture is to apply a machine learning method based on the distributed and parallel environment to the new system. Additionally, a scheduler to control the whole knowledge base construction process from raw data crawling to knowledge provision to service is introduced. The new system is expected to be faster and more accurate than the existing system.

2 Related Work

It is generally known that machine learning methods are more preferred recently than handcrafted rules on which early studies were mostly based [1]. If the training data to learn systems is large enough to guarantee the quality of extraction, machine learning methods are also more accurate than other methods.

SystemML is a system package developed by IBM to enable a variety of machine learning algorithms to be executed in a MapReduce based distributed processing environment [2]. SystemML is important in that the existing knowledge extraction algorithms are implemented to be driven in the distributed processing environment in order to process big data. Mahout[1] is also intended to provide a variety of machine learning algorithms through Mahout as a library. The idea is that the library is to extend the library effectively in a cloud environment by using the Apache Hadoop to solve the issue of processing time taken to learn large data set which is one of disadvantages of existing machine learning algorithms. Exemplary open sources include Lucene in charge of pre-processing of machine learning, Hadoop which enables the machine learning algorithm to be executed in the distributed processing environment, and Hama which enables MapReduce to be effectively used.

From the related work, the implications are mainly two. First, machine learning methods with large training data are generally more precise than rule based methods. Second, machine learning methods can be executed in a MapReduce based distributed processing environment.

3 Analysis of the Existing System

Figure 1 shows the process of knowledge extraction used for the InSciTe Adaptive service. InSciTe Adaptive service is a user adaptive intelligent service to support making decisions on things related to technologies or products. In order to supply the knowledge base for the service, the existing system applies a rule-based information extraction method (step 4), and the process contains some post processing tasks (step 5, 6, 7) to guarantee the quality of data. After extraction, reasoning (step 8, 9) is made to make semantic triples to extend the outputs of extraction. The system addresses 5.3 million web articles, 9.8 million papers, and 7.6 million patents, and it makes about 500 million semantic triples, which takes approximately 5 days.

[1] http://mahout.apache.org/

The existing system has mainly two drawbacks. One is that it takes more time to extract knowledge and build the knowledge base than it is expected because the system is implemented to run on single machine. The systems which operate on single server are not able to deal with large amounts of data due to physical resource limitations. It influences the accuracy of information extraction result. The other is that it adopts a rule-based knowledge extraction method which is typically domain dependent and requires high cost with significant amount of manual efforts [3].

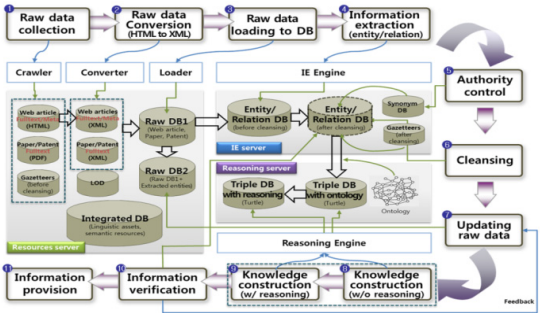


Fig. 1. The Existing Process and Architecture

4 Proposed System

We propose a new knowledge extraction system whose process shown in Figure 2. Considering the implications of related works, unlike the existing system, a machine learning method is hired and the system is executed on a distributed and parallel environment: MapReduce framework and Hadoop file system are applied to the new system. The process starts with crawling raw data such as web articles, papers, and patents. After filtering and converting them, it does preprocessing tasks on input data such as parsing and PoS (Part of Speech) tagging. At the same time, it builds extraction model using a machine learning method. With the extraction model, it then extracts knowledge and optimizes the module to gain high quality of output data. Finally, it builds triple store with reasoning. All tasks except for crawling are supported by a distributed and parallel method.

The proposed architecture is based on distributed and parallel environment. Figure 3 shows each part of proposed system. It is composed of three parts; Data collection (left side), knowledge extraction based on MapReduce and Hadoop (right side), and job management (top side). On each slaver server, the modules for the tasks such as preprocessing, information extraction, triple store construction, and reasoning are installed and executed. The master server has a knowledge extraction management module for the task management of each slave server, an input document management module for management of the first entered data, and an output document management module for the management of the final output data. The MapReduce framework is attracting attention, as a cluster consisting of a large number of low-cost server processes data in a distributed and parallel method [4].

One of the expected issues for the new system is related to the job management, especially to the job scheduler. The system should be automatically executed according to the sequential steps to reduce the idle time of the system. For this, the scheduler is designed to coordinate all tasks and makes them processed in order

without the system waiting. Another purpose of the scheduler is that even though the system fails to run modules and can't notice the state to the scheduler, the scheduler should address the situation and make jobs go forward. For solving this issue, we use the log table method. It records the state in the log table after a task is fully done. When errors occur, the state is not recorded in the log table so that the scheduler makes the current module restart by checking the log table.

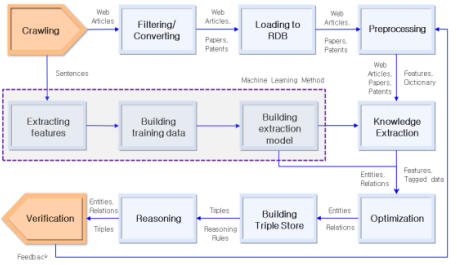


Fig. 2. Proposed process

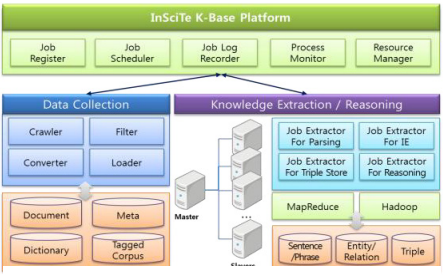


Fig. 3. Proposed architecture

5 Comparison of Two System

The comparison is divided into three categories; differences in process and architecture, input and output data, and processing time. Differences in process and architecture are extraction method applied, processing environment, and execution (Table 1). The existing system uses a single machine based rule and dictionary method which is applied only to the knowledge extraction task. On the other hand, all tasks except for crawling are executed by a cluster based machine learning method in the new system. It is also automatically driven by a scheduler. The volume of raw data is also larger in the new one than in the existing one as well as the output data (triples) of the new system are more.

Table 1. Comparison between the existing system and the new system

Criteria	Existing system	New system
Extraction Method	Rule and dictionary base	Machine learning (80,000 sentences for training)
Processing Environment	Single machine	Cluster (Hadoop and MapReduce)
Execution	Manual execution	Automatic execution by scheduler
Volume of input data	5.3 million web articles, 9.8 million papers, 7.6 million patents	Over 10 million web articles, over 10 million papers, over 8 million patents
Volume of output data	500 million triples	Approximately 1 billion triples
Processing time	5 days	Less than 3 days
Evaluation (F1 score)	78% (600 web documents for test)	Over 70% (5,000 web documents for test)

The new system is expected to be faster and more accurate than the existing one because of a machine learning method based on a cluster environment. From the previous study done by J. H. Um et al. [4], the performance of a knowledge extraction module with a cluster based machine learning method is nearly twice as faster as the same existing system. Based on the previous result, the new system is expected to have less processing time by approximately 2 days than the existing system (Table 1). In terms of accuracy, the result of the existing system is 79% in F1 score with 600 test documents [5]. Unlike the existing system, the new system will be used for more practical purpose so that more data for test is needed. 100,000 sentences is about 5,000 web documents if it is assumed that one web document has in average about 20 sentences. Because the evaluation with about 5,000 documents in research is beyond evaluation: it is almost the same as the job in real situation. Therefore, the goal of evaluation with 100,000 random sentences is 70% in F1 score, but it is more valuable goal than 78% with 600 web documents for test.

6 Conclusion

We have reviewed the existing system which is implemented to use single machine based rule and dictionary method to build knowledge base. It is lacking of the performance in terms of velocity and accuracy. Compared with the existing system, the proposed system is designed to equip with the distributed and parallel computing technology using the MapReduce and Hadoop. The system is also based on a machine learning method. Even though the proposed system processes more input data than the existing system, it is expected to be faster in building knowledge base. The accuracy is also more valuable because much more documents for test are used.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigations 30, 3–26 (2007)
2. Ghoting, A., Krishnamurthy, R., Pednault, E., Reinwald, B., Sindhwani, V., Tatikonda, S., Tian, Y., Vaithyanathan, S.: SystemML: Declarative machine learning on MapReduce. In: Proceedings of the 2011 IEEE 27th International Conference on Data Engineering, pp. 231–242 (April 2011)
3. Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., Vaithyanathan, S.: Domain adaptation of rule-based annotators for named-entity recognition tasks. In: EMNLP, pp. 1002–1012 (2010)
4. Um, J.H., Shin, S., Choi, Y.S., Jeong, C.H., Song, S.K., Choi, S.P., Jung, H.: A Knowledge Extraction System using the MapReduce Framework for Massive Amounts of Technical Data. In: The 2nd Joint International Semantic Technology Conference (December 2012)
5. Chun, H.W., Jeong, C.H., Shin, S., Seo, D., Hwang, M.N., Jang, H.J., Park, J.W.: Information Extraction for Technology Trend Analysis. AISS: Advances in Information Sciences and Service Sciences (accepted)

A Novel Anomaly Detection System Based on HFR-MLR Method

Eunhye Kim[1] and Sehun Kim[2]

[1] IT Convergence Technology Research Division, ETRI, South Korea
eunhye@etri.re.kr
[2] Internet Security Lab., KAIST, South Korea
shkim@kaist.ac.kr

Abstract. Reducing the data space and then classifying anomalies based on the reduced feature space is vital to real-time intrusion detection. In this study, a novel framework is developed for logistic regression-based anomaly detection and hierarchical feature reduction (HFR) to preprocess network traffic data before detection model training. The proposed dimensionality reduction algorithm optimally excludes the redundancy of features by considering the similarity of feature responses through a clustering analysis based on the feature space reduced by factor analysis, thus helping to rank the importance of input features (essential, secondary and insignificant) with low time complexity. Classification of anomalies over the reduced feature space is based on a multinomial logistic regression (MLR) model to detect multi-category attacks as an outcome with the goal of reinforcing detection efficiency. The proposed system not only achieves a significant detection performance, but also enables fast detection of multi-category attacks.

Keywords: Anomaly detection, Dimensionality reduction, Hierarchical clustering, Multinomial logistic regression.

1 Introduction

As the potential damage caused by malicious network activities has become more serious, the need to defend against these threats has increased significantly. The network intrusion detection system (NIDS), as a vital system in the network security infrastructure, aims to detect attacks quickly and accurately; its role is becoming more important. To achieve this objective, previously observed attack patterns need to be analyzed and profiled so that criteria for what constitutes normal traffic or an attack can be determined and applied to newly captured patterns for intrusion detection. In the detection approaches of NIDS, many studies have applied data mining techniques such as a support vector machine (SVM) and neural networks [1-2].

Although the techniques applied in previous works have shown good results in terms of data classification, they are not favorable for large-scale datasets because the training complexity is very much dependent on the amount of data in the training set. Especially, some data features in the classifiers used in NIDS may be redundant or

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_43,

may contribute little to the detection process. Extraneous features and the complex relationships that exist among the features can make it harder to detect suspicious behavior patterns and can increase the computation time. Therefore, through feature dimensionality reduction, NIDS must reduce the amount of data to be processed for computationally efficient and effective detection.

This study proposes a multinomial logistic regression (MLR)-based network anomaly detection system based on hierarchical feature reduction (HFR) to preprocess network traffic data before detection model training. The proposed HFR algorithm optimally excludes the redundancy of features by considering the similarity of feature responses through a clustering analysis based on the feature space reduced by factor analysis. The performance of the proposed method is evaluated using different data sets reduced by the ranking of the importance of input features. Classification of intrusions over the reduced feature space was based on the MLR model, a method well suited for analyzing multi-type outcomes with high speed in learning techniques. Our classification model was developed for the detection of multi-category attacks as an outcome to reinforce detection efficiency, unlike previous studies that were focused on a binary outcome (e.g., normal or abnormal). The experiment with the NSL-KDD dataset showed a significant detection rate through a good subset of features with a significant improvement in speed.

This paper is organized as follows. In Section 2, several examples of related work are reviewed. The proposed algorithm is then described in Section 3. Section 4 gives details of the experiments as well as the results. The study is concluded with a summary and plans for future research in Section 5.

2 Related Work

Anomaly detection depends on the idea that the characteristics of normal behavior can be distinguished from those of abnormal behavior. Statistical modeling remains the most common approach to anomaly intrusion detection; this method includes cluster analysis, Bayesian analysis, principal component analysis, and the fuzzy inference approach. Leung et al. [3] carried out research based on density and a grid-based clustering method for anomaly detection. Chan et al. [4] investigated both the distance and density of clusters and found that attacks were often in outlying clusters with statistically low or high densities. Valdes et al. [5] employed naive Bayesian networks to perform intrusion detection on traffic bursts. Xu et al. [6] used continuous time Bayesian networks and avoided specifying a fixed update interval common to discrete-time models. Huang et al. [7] presented a simple algorithmic framework for network-wide anomaly detection that relies on distributed tracking combined with approximate PCA. Toosi et al. [8] combined a neuro-fuzzy network, the fuzzy inference approach, and genetic algorithms to design an intrusion detection system.

Most previous studies were conducted based on all possible independent variables. Unnecessary variables can create bias and lead the model either to overestimate or underestimate the detecting values. In this study, in order to reduce the amount of training data, the HFR method was developed using unsupervised data mining

techniques, and applied before MLR training. Also, our classification model was developed for the detection of multi-category attacks as an outcome to reinforce detection efficiency, unlike previous studies that were focused on a binary outcome.

3 Proposed Framework

The proposed framework consists of three main phases. In the first phase, the feature redundancy can be reduced by considering the similarity of variable-responses to the training data set through clustering analysis. The proposed scheme can hierarchically reduce the features, thus helping to rank the importance of input features. Then, in the second phase an anomaly detection model using MLR is constructed with the reduced training dataset resulted from the feature reduction algorithm. As a result of the model, the odds ratios provide an estimate of the likelihood of being identified as an anomaly. In the third phase, test data are used to detect anomalies according to attack types based on the developed MLR model. The performance of our anomaly detection model is evaluated using a cross validation testing concept.

3.1 Proposed Hierarchical Feature Reduction

Feature reduction involves processes of determining the evidence that can be taken from the raw data that is most useful for analysis. To exclude the redundancy of features and to improve the performance of classification, statistical techniques are used, including factor analysis which is one of the most widely used dimensionality reduction techniques and hierarchical clustering which does not require predetermined numbers of groups and has the advantage of low time complexity.

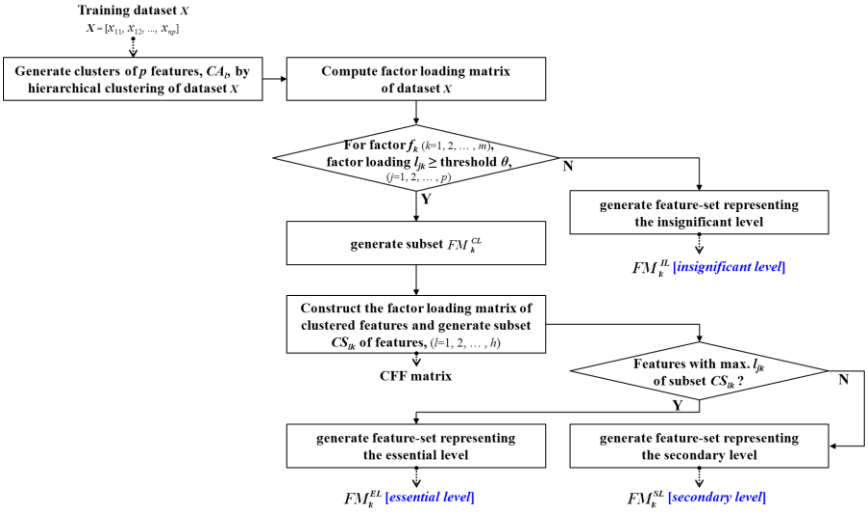


Fig. 1. Proposed HFR procedure

In an application of factor analysis, if feature dimensionality reduction is based on only the degree of contribution induced from observing which variables are most heavily loaded on certain factors, the selected features may be redundant as the information that they include is contained in other features. This redundancy can be reduced by considering the similarity of variable-responses to the training data set through clustering analysis.

Therefore, a hierarchical feature dimensionality reduction algorithm is proposed in which the factor analysis and hierarchical clustering are combined. In the proposed algorithm, hierarchical clustering is initially applied and factor analysis is then applied to the training data set, as shown in Fig. 1. Based on h clusters of features through hierarchical clustering, features are extracted in which the factor loadings are higher than a certain threshold (subset CS_{lk} in Fig. 1). The redundancy of features with a high value of factor loadings is then reduced if they are in the same cluster, which organizes a good subset (subset FM_k^{EL}) of features critical to the performance of classifiers. The strongest point of the proposed feature reduction scheme is that this method can hierarchically reduce the features, thus helping to rank the importance of the input features (essential, secondary and insignificant) with low time complexity. Using far fewer instances, the proposed method can produce high quality datasets that sufficiently represent all of the instances in the original dataset. The clustered feature-factor (CFF) matrix generating subset of the significant features is shown in Fig. 2.

		FACTOR												Subset No.
Cluster No.	Feature Label	1	2	3	4	5	6	7	8	9	10	11	12	
CA 1	FL_A	.0464	.2902	.0088	.5173	.0491	-.0692	-.0720	.1094	-.1855	-.0371	.1889	.3709	
	FL_AD	-.0131	.0015	.0117	.1788	-.8820	.0009	-.0124	-.0113	-.0242	-.0282	.0438	-.0071	*CS1,5*
	FL_AI	-.0449	-.0067	.0379	.7953	-.0940	.0226	-.0114	-.0239	-.0728	-.0136	-.0058	.0571	*CS1,4*
	FL_AJ	.0444	-.0279	.0494	.6782	.0207	.0401	-.1267	-.1935	.2833	-.0143	.2356	-.0509	*CS1,4*
CA 2	FL_B	-.0820	-.0252	.7639	.2710	-.2616	-.0393	-.0591	-.1033	.0045	-.0387	.0501	-.0257	*CS2,3*
	FL_C	-.0914	.0511	.7243	.1332	.0603	-.0010	-.0639	.0948	-.1125	-.0264	-.3450	.0679	*CS2,3*
	FL_W	-.0605	-.0068	.8537	-.1957	.0912	-.0145	.0125	-.0149	-.0294	-.0064	.1633	-.0143	*CS2,3*
	FL_X	-.0762	-.0081	.8840	-.2102	.1346	-.0249	.0037	-.0318	-.0308	-.0102	.1454	-.0103	*CS2,3*
	FL_AF	-.2046	-.0138	.4171	-.0764	-.2146	-.0681	.0445	.1982	-.4915	-.0147	.0237	.0115	
CA 3	FL_D	-.9610	-.0041	.0190	.0069	-.0047	-.1516	.0100	-.0137	-.0493	.0044	.0123	-.0121	*CS3,1*
	FL_L	-.3389	.0388	-.7414	-.3204	.1770	-.0319	.0575	.0433	-.0453	-.0202	-.0081	.0296	*CS3,3*
CA 4	FL_E	-.0121	-.0184	-.0256	.1043	.0428	.0395	-.0082	-.0763	-.0496	.7551	-.0058	-.0490	*CS4,10*
	FL_F	.0021	.0615	-.0179	-.0325	.0067	-.0517	-.0173	.2243	-.0221	.6678	.0838	.0293	*CS4,10*
CA 5	FL_G	-.0563	-.0027	.0638	-.0662	-.0613	.0216	.0196	.2986	.5855	-.1017	-.0278	-.0025	
CA 6	FL_J	-.0086	.0009	-.0374	.0585	.0289	-.0013	.9265	-.0043	-.0152	-.0227	.0426	.0384	*CS6,7*
	FL_V	-.0113	-.0036	-.0492	.0610	.0016	-.0094	.9379	-.0206	-.0067	-.0066	.0234	-.0243	*CS6,7*
CA 7	FL_K	.0135	.0230	-.0054	.0254	.0104	.0026	-.0050	.0468	-.0068	-.0025	-.0053	.4479	
CA 8	FL_M	.0010	.9472	.0037	.0217	.0031	-.0237	-.0008	.0151	.0064	.0203	-.0010	-.0039	*CS8,2*
	FL_N	.0030	.7099	-.0075	.0130	-.0004	.0501	.0089	-.0183	.0393	.0042	.0230	.0166	*CS8,2*
	FL_O	.0045	.8814	-.0003	.0349	.0017	.0643	-.0003	.0186	.0091	.0108	.0128	.0930	*CS8,2*
	FL_P	.0006	.9473	.0026	.0221	.0022	-.0232	-.0012	.0112	.0075	.0203	-.0010	-.0089	*CS8,2*
	FL_S	-.0028	.7781	-.0162	.0046	-.0041	-.0312	-.0067	.0442	-.0281	-.0023	-.0333	-.0100	*CS8,2*
CA 9	FL_Q	.0222	.0190	-.0397	.1024	.0953	-.0044	.0110	.2181	-.1749	-.1654	.1709	.5928	
CA 10	FL_R	.0313	.0157	-.0610	.1411	.1527	-.0485	-.0482	.2236	-.2575	-.2147	.2317	-.6332	*CS10,12*
CA 11	FL_Y	.0108	.0161	-.0261	.0106	-.0201	.9477	-.0075	.1545	.0119	-.0068	-.0039	.0044	*CS11,6*
	FL_Z	.0218	.0238	-.0276	.0082	.0076	.9505	-.0037	.1234	.0219	-.0032	.0057	.0138	*CS11,6*
	FL_AL	-.0085	.0100	-.0293	.0928	-.0003	.1235	-.0131	.7730	.0156	.0359	-.0199	.0637	*CS11,8*
	FL_AM	-.0055	.0459	-.0095	.0401	.0085	.1362	-.0124	.7773	.1152	.1005	.0177	.0752	*CS11,8*
CA 12	FL_AA	.9751	.0060	-.0212	-.0030	.0096	-.0393	-.0106	-.0067	.0395	-.0043	-.0137	.0182	*CS12,1*
	FL_AB	.9706	.0068	-.0222	-.0043	-.0005	-.0469	-.0086	-.0051	.0295	-.0053	-.0163	.0200	*CS12,1*
	FL_AK	.3107	.0407	-.1161	.0942	.0729	-.0168	-.0254	-.0011	.6668	.0003	.0583	-.0360	*CS12,9*
	FL_AN	.9199	-.0080	-.0290	.0037	.0208	-.0116	.0058	-.0232	.0779	-.0001	-.0300	-.0165	*CS12,1*
	FL_AO	.9274	.0006	-.0335	-.0084	.0105	-.0050	-.0016	-.0071	.0867	-.0027	.0000	.0113	*CS12,1*
CA 13	FL_AC	.0133	.0039	-.0384	-.1984	.8903	-.0145	.0242	-.0008	.0429	.0317	-.0528	.0141	*CS13,5*
	FL_AG	-.0320	-.0362	.0147	-.7816	.2319	-.0041	-.2013	-.1875	-.0912	-.0924	.1563	.0324	*CS13,4*
	FL_AH	.0539	-.0298	.0018	-.7401	.3065	.0137	-.2099	-.2436	.0657	-.0984	.1456	.0032	*CS13,4*
CA 14	FL_AE	.0693	-.0062	-.1097	.0007	.0892	-.0036	-.0633	-.0034	-.0109	-.0824	-.8500	-.0253	*CS14,11*

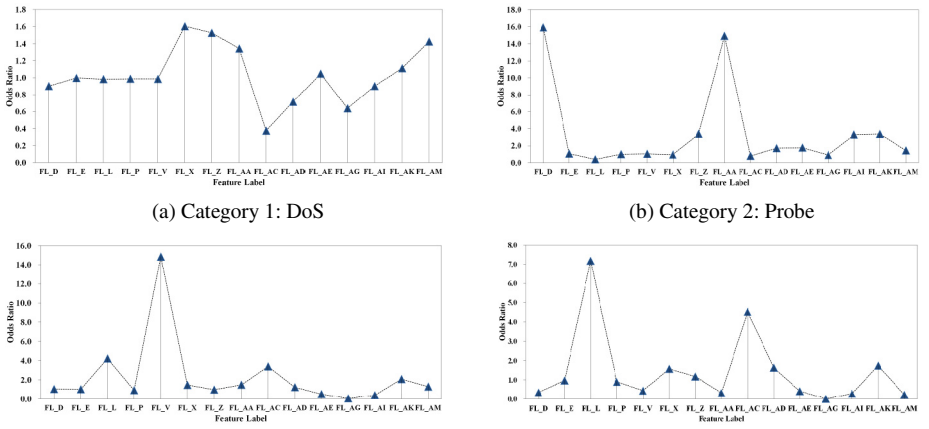
Fig. 2. Clustered feature-factor matrix of the training dataset

3.2 Anomaly Detection Method

The proposed anomaly detection method uses an MLR to build a classifier model. Unlike a binary logistic model, in which a dependent variable has only a binary choice, the dependent variable in the MLR model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category [9]. This study used '0' (normal) as the reference category. Suppose Y_i is the dependent variable with five categories for individual connection i; the probability of being in category m can be represented with the chosen reference category:

$$\log \frac{P(Y_i = m)}{P(Y_i = 0)} = \alpha_m + \sum_{k=1}^{p} \beta_{mk} x_{ik} = Z_{mi} \tag{1}$$

where m='1' [DoS], '2' [Probe], '3' [R2L], and '4' [U2R]. Our MLR modeling is performed with a significance threshold of 0.05 for adding variables and an insignificance threshold of 0.1 for removing variables, yielding a set of variables that are associated with the outcome in a statistically significant way. The final MLR model calculates the predicted probabilities of being in the outcome category for each connection record; the classification of the unordered set {0, 1, 2, 3, 4} is conducted on the basis of that probability. The odds ratio of the proposed MLR model, consisting of the essential level variables, for detecting each attack relative to the normal category is shown in Fig. 3.



(a) Category 1: DoS

(b) Category 2: Probe

(c) Category 3: R2L

(d) Category 4: U2R

Fig. 3. Odds ratio of the essential features by attack category

4 Experiments and Results

The data used for testing is NSL-KDD, which is a new dataset for the evaluation of studies in network intrusion detection systems [10]. Each NSL-KDD connection record contains 41 features (e.g., protocol type, service, and flag) and is labeled as

either normal or an attack, with one specific attack type. The attacks fall into one of the four categories: DoS, Probe, R2L, and U2R. The NSL-KDD training set contains a total of 22 training attack types, with an additional 17 types in the test set. In the experiments, the dataset was partitioned into subsets. The training set contained 125,973 records and five evaluation sets were comprised of 111,630 records.

We compared the evaluation results of models using a selected feature-set in the essential level with those using essential and secondary level features by the proposed feature reduction algorithm. As can be seen in Fig. 4, the classification rates using the feature-set in the essential level are comparable to those using the essential and secondary levels, except for the case of the Probe class. Both sets of performance results show difficulties in detecting R2L attacks, which are embedded in the data packets themselves and do not form a sequential pattern. These were assigned to incorrect classes and lowered the detection rate. And, with the too small number of instances of U2R attacks in the NSL-KDD dataset, both models of reduced features provided relatively low performance for the U2R class. However, compared with the results of the performance with essential and secondary level features, test numbers 1, 6, 10, 11, 13, and 14 showed higher detection rates with lower false alarm rates, as shown in Fig. 4. It can be said that the proposed HFR method achieves significant detection rates that demonstrate the possibility of successfully detecting attacks with a significant improvement in speed by using only a half percent of the comparison feature-set and 39.0% as compared with the full feature-set. Our method also improved detection times by 23.8% compared to those including the secondary level features.

Experiments were also attempted to evaluate the performance of our anomaly detection scheme compared with that of several other methods; results are shown in Table 1. It can be stated that all the algorithms tested on the KDD data set offered an acceptable level of detection performance for Normal, DoS and Probe classes; they did not have good performance on R2L and U2R attacks. The SVM with BIRCH clustering [11] and ESC-IDS [8] showed the best detection rate for the DoS attack, and Multi-classifier [13] showed good detection rate for the Probe and U2R attacks. Works by Xuren et al. [12] provided the best performance for the normal class. Our proposed method demonstrated a better detection rate for R2L attacks and provided comparable performance for Probe and U2R attacks.

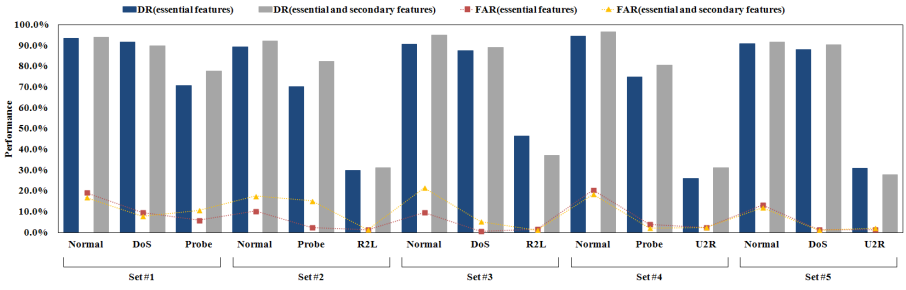


Fig. 4. Comparison of performance results: DR and FAR

Table 1. Comparison of performance results with other works

Method	Normal	DoS	Probe	R2L	U2R
Proposed method with the essential features	0.917	0.890	0.719	0.381	0.284
Proposed method with the essential and secondary features	0.937	0.897	0.802	0.342	0.295
SVM with BIRCH clustering (Horng et al., 2011)	0.993	0.995	0.975	0.288	0.197
ESC-IDS (Toosi et al., 2007)	0.982	0.995	0.841	0.315	0.141
Association rule (Xuren et al., 2006)	0.995	0.968	0.749	0.079	0.038
Multi-classifier (Sabhnani et al., 2003)	n/r	0.973	0.887	0.096	0.298

5 Conclusion

In this paper, an HFR method that combines hierarchical clustering and factor analysis was introduced; an anomaly detection approach based on an MLR was presented. Experimental results show that the proposed system could achieve a significant detection performance by using only a half percent of the comparison feature-set and 39.0% as compared with the full features. Our method also improved detection times by 23.8% compared to those including the secondary level features. Therefore, it can be concluded that our approach can efficiently reduce the features that are redundant or that hinder the process of detecting intrusions. The proposed method enabled reinforcing detection efficiency by the detection of multi-category attacks as an outcome. Future research will include the integration of various probabilistic techniques to achieve better detection performance and the accuracy of predictions.

References

1. Kabiri, P., Ghorbani, A.A.: Research on Intrusion Detection and Response: a Survey. Int. J. Netw. Sec. 1, 84–102 (2005)
2. Lazarevic, A., Ozgur, A., Ertoz, L., Srivastava, J., Kumar, V.: A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In: SIAM International Conference (2003)
3. Leung, K., Leckie, C.: Unsupervised Anomaly Detection in Network Intrusion Detection. In: Australasian Computer Science Conference (2005)
4. Chan, P.K., Mahoney, M.V., Arshad, M.H.: Learning Rules and Clusters for Anomaly Detection in Network Traffic. In: Managing Cyber Threats: Issues, Approaches and Challenges, pp. 81–99. Springer (2005)
5. Valdes, A., Skinner, K.: Adaptive Model-based Monitoring for Cyber Attack Detection. In: Recent Advances in Intrusion Detection Toulouse, pp. 80–92 (2000)
6. Xu, J., Shelton, C.R.: Intrusion Detection using Continuous Time Bayesian Networks. J. Art. Int. Res. 39, 745–774 (2010)

7. Huang, L., Nguyen, X., Garofalakis, M., Jordan, M.I., Joseph, A., Taft, N.: In-Network PCA and Anomaly Detection. In: Neural Information Processing Systems, pp. 617–624 (2006)
8. Toosi, A.N., Kahani, M.: A New Approach to Intrusion Detection based on an Evolutionary Soft Computing Model using Neuro-Fuzzy Classifiers. Com. Comm. 30, 2201–2212 (2007)
9. McFadden, D.: Conditional LogitAnalysis of Qualitative Choice Behavior. Frontiers in Econometrics, 105–142 (1974)
10. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.: A Detailed Analysis of the KDD CUP 99 Data Set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications (2009)
11. Horng, S.J., Su, M.Y., Chen, Y.H., Kao, T.W., Chen, R.J., Lai, J.L., Perkasa, C.D.: A Novel Intrusion Detection System based on Hierarchical Clustering and Support Vector Machines. Exp. Sys. W. Appl. 38, 306–313 (2011)
12. Xuren, W., Famei, H., Rongsheng, X.: Modeling Intrusion Detection System by Discovering Association Rule in Rough Set Theory Framework. In: International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (2006)
13. Sabhnani, M.R., Serpen, G.: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. In: International Conference on Machine Learning: Models, Technologies, and Applications, pp. 209–215 (2003)

Knowledge Discovery and Integration: A Case Study of Housing Planning Support System

Junyoung Choi[1], Daesung Lee[2], and Hanmin Jung[3]

[1] Spatial Information Office, Korea Land and Housing Corporation, Korea
junyoung@lh.or.kr
[2] Dept. of Computer Engineering, Catholic University of Pusan, Korea
dslee@cup.ac.kr
[3] Korea Institute of Science and Technology Information, Korea
jhm@kisti.re.kr

Abstract. Different information related to knowledge should be collected and filtered by system in order to discover and integrate knowledge in multiple knowledge representation environment. Especially, various information exist in multiple forms and in separate spaces, such as geographic information, housing statistics, and policy statements, in housing planning doamin. Housing policy support system is needed to make a proper spatial decision for supply housing when housing demand is occured. Thus, this paper proposes a conceptual design of system for knowledge discovery and integration in housing planning support. This system uses GIS based Housing Demand and Supply Mapping Model (HDSMM) to support policy decision using housing statistics and geographic information and to be composed of multi-dimension analysis, policy monitoring functionalities.

Keywords: knowledge discovery, knowledge integration, housing supply, housing demand, GIS.

1 Introduction

Knowledge discovery and integration is the processes of extracting and synthesizing specific knowledge at multiple knowledge models and environments. Knowledge integration is related to the understanding of a specific subject from various perspectives as well as merging various information with different schemas [1]. Knowledge discovery is the creation of knowledge from various sources such as database, xml documents, text and images [2].

There exist various information related to housing policy, such as geographic information, housing statistics, and policy statements, in internet. These information should be collected and analyzed for knowledge extraction and integration related to housing planning. However, these were represented in different schemas and representation models. Also, various perspectives should be considered for housing planning support system. They are rationalization of overlapped regulation on housing redevelopment and reconstruction, deregulation on multi-households and multiplex

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_44,

housings construction and so forth [3]. This requires knowledge discovery and integration system that analyze and provide conditions of demand supply and necessitates construction of Housing Demand and Supply Mapping Model (HDSMM) which provides analytic information about housing demand and supply along with geographic information on administrative boundaries, urban zones of life, metropolitan traffic networks, locations of housing construction, public and private development project zones and so forth [4,5,6].

This study summed up the knowledge about housing demand and supply by connecting and integrating with systems like electronic Architecture administration Information System, Housing supply statistics Information System, land for housings information system and constructed an integrated housing database on that linkage. This paper is organized as follow: In Section 2, background knowledge to understand our proposal is explained. Section 3 draws out design of housing planning support system for knowledge discovery and integration, and finally result is summarized in Section 4.

2 Background Knowledge

2.1 Knowledge Discovery and Integration

Knowledge discovery is to extract knowledge from various sources (databases, XML documents, text, documents, images). The discovering knowledge needs to be able to read and interpreted by a machine, and should represent knowledge in a manner through reasoning. Its main purpose is that the extraction result should cover beyond the creation of structured information in a relational schema although it is similar to Information Extraction in processing method. It needs that existing formal knowledge can be reused, or a schema based on the source data can be generated [2].

Knowledge integration is the process that merges different knowledge models into a common model. It focuses on the understanding of a given subject from different perspectives compared to information integration. It has also been researched as the process of combining new information into existing knowledge with an interdisciplinary approach. This process includes defining the interaction between the new information and the existing knowledge, the modification of existing knowledge to new information orientationaly, and how the new information should be modified in light of the existing knowledge [1].

2.2 Housing Demand and Supply

Housing demand and supply can be defined either respectively or in a pair [7]. In a respective definition, housing demand may be divided into housing needs and housing demand. Housing needs refers to households lacking their own housing or living in unsuitable housing, whereas housing demand is quantity or quality of housing which households afford to buy or rent in the market. On the other hand, housing supply is to provide physical or non-physical services by securing housing sites and it takes

long time to supply housing. This is why housing supply, in the Korean housing market, is classified into housing permit and approval, housing construction start and completion, housing sale and housing destruction by housing construction life cycle.

If defined in a pair, housing demand and supply can be understood in the quantitative perspective. This is the concept derived by comparing aggregate housing demand with aggregate housing supply, or new housing demand with new housing supply. In this perspective, estimating demand and supply by quantity is advantageous to policy but disaggregated demand is needed to cope with local housing market. Summing up the above discussion, while housing supply is defined in the function of housing permit and approval, housing construction start and completion, housing demand is defined in the function of population, housing price, incomes. The operational definition of housing demand and supply is the gap between housing demand and supply.

HDSMM is a housing supply monitoring system through which departmentalized information of housing demand and housing supply can be identified on the map at one glance. However, as HDSMM can be understood in multiple ways combined with housing demand, supply and map, so various concepts are possible according to housing studies, information systems, and geographic information science. Thus, HDSMM will be conceptualized in terms of data and analysis model perspective, information system perspective and service perspective for clear definition.

3 Housing Planning Support System for Knowledge Discovery and Integration

3.1 Housing Planning Support System Design Elements

Design elements can be derived from each perspective according to a model for housing demand and supply mapping. We derived four elements from each perspective: housing demand and supply, GIS, On-Line Analytical Processing (OLAP), and monitoring, as shown in Table 1. First, a housing demand and supply model can be constructed after factors affecting housing demand and supply are derived. It is constructed as a database defining factors as properties.

Table 1. Housing Planning Support System Design Elements

Perspective	Needed services	Design Elements
Data analysis model	Accumulate housing information Analyze the regional demand-supply	Housing demand, Housing supply, GIS
Information System	Process housing information Provide tailored housing demand-supply information	OLAP, GIS
Service	Browse the housing information in the map Diagnose the regional housing status	GIS, Monitoring

The housing demand and supply model is developed to forecast short and long-term demand, and supply, for housing policy support. Second, GIS is used to construct geographical information and to develop geographic analysis methods. It develops methods to indicate housing demand and supply information with maps and to conduct geographic analysis, such as demand and supply of housing in cities, available land for housing, and station-influenced areas. Third, OLAP is conducted via building a housing data mart and data warehouse. Fourth, the housing market is monitored to check the status of demand and supply and the effectiveness of policies.

Housing planning support system is composed of basic statistics, housing demand amounts, housing supply amounts, and housing demand and supply. Basic statistics include population size, number of households, number of families, housing supply, housing inventory, and housing destruction. These are used for measuring housing demand and housing supply amounts.

- **Basic Statistics**

Korean population and housing census (hereinafter defined as 'census') data are used as reference values for basic statistics. Data from the census year are directly applied as reference values for the basic statistics. However, data from other years was indirectly estimated using data from the census year and other data.

- **Housing Demand**

Housing demand usually uses estimated values for mid and long-term demand. The modified model of Mankiw-Weil is applied for estimating. This model has demographical factors, such as population and household changes in the original model, as major variables. This study applied data from previous studies, such as comprehensive housing plans, and used the M-W model and a modified M-W model [8].

- **Housing Supply**

Housing supply amounts are dealt as units corresponding to administrative boundaries, such as city limits, guns (counties) and gus (boroughs) for each housing construction step. It has attributes of a new housing supply, housing stock, and destruction. Information related to housing supply amounts differ, depending on whether they are apartment houses or detached houses. While information on apartment houses can be obtained from census data, information on the distribution and entrances of residences cannot be created for detached houses. Thus, the amount of entrances in detached houses needs to be estimated from the amount of housing completed on a case-by-case basis. Census data is applied to estimate housing stock for the base year, while new housing supply and destruction are added and subtracted for other years.

- **Housing Demand and Supply Amount**

Housing demand and supply is estimated as the difference between the amount of demand and supply, or the amount of new housing demand and new housing supply. The housing demand and supply is identified by enabling accumulation by year and

location. Design elements can be derived from each perspectives presented in the previous Section. We derived elements from each perspective, and then design elements of HDSMM are classified as shown in Table 2.

Table 2. Design Elements

Perspective	Design Elements
Data analysis model	Integrated housing database (DB), spatial analysis, monitoring
Information System	Housing Information System, spatial analysis, DB
Service	GIS based services, OLAP, DB

4 Conclusion

We proposed Housing Planning Support System for knowledge discovery and integration and its concepts of each component are derived to make the mapping model which is used to recognize the regional state of demand and supply to offer various kinds of data about the housing supply and demand. To achieve this aim, various kinds of data for housing supply and demand are used. In the future study, when real data for housing transactions are acquired, it will be possible to better understand the regional housing market by monitoring and simulating housing demand-supply more precisely.

References

1. http://en.wikipedia.org/wiki/Knowledge_integration
2. http://en.wikipedia.org/wiki/Knowledge_discovery
3. Gang, Y., Lee, J., Park, M.: Guidelines for the Improvement of Accuracy on Building related Registers Information. J. Korea Spatial Information System Society 8(3), 15–26 (2006)
4. Hanushek, A.E., Quigley, M.J.: What is the Price Elasticity of Housing Demand? The Review of Economics and Statistics 62(3), 449–454 (1980)
5. Korea National Housing Corporation: Development of Analysis Model and Housing Indicator using Housing Supply Statistics (2008)
6. Saiz, A.: The Geographic Determinants of Housing Supply. The Quarterly J. Economics 125, 1253–1296 (2010)
7. Green, K.R., Malpezzi, S., Mayo, K.S.: Metropolitan-Specific Estimates of the Price Elasticity of Supply of Housing, and Their Sources. The American Economic Review 95(2), 334–339 (2005)
8. Bramley, G., Pawson, H.: Low demand for housing: incidence, causes and UK national policy implications. Urban Studies 39(3), 393–422 (2002)

Performance Analysis of MapReduce-Based Distributed Systems for Iterative Data Processing Applications

Min Yoon[1], Hyeong-il Kim[1], Dong Hoon Choi[2], Heeseung Jo[1], and Jae-woo Chang[1,*]

[1] Dept. of Computer Engineering, Chonbuk National University, Jeonju, Republic of Korea
[2] Korea Institute of Science and Technology Information, Daejeon, Republic of Korea
{myoon,melipion,heeseung,jwchang}@jbnu.ac.kr, choid@kisti.re.kr

Abstract. Recently, research on big data has been actively made because big data are generated in various scientific applications, such as biology and astronomy. Therefore, distributed data processing techniques have been studied to manage the big data in large number servers. Meanwhile, some scientific applications like genome data analysis require loop control in analyzing big data using a MapReduce framework. In this paper, we first describe the existing MapReduce-based distributed systems which support iterative data processing. In addition, we do the performance analysis of the existing distributed systems in terms of execution time for various scientific applications which require iterative data processing. Finally, based on the performance analysis, we discuss some requirements for a new MapReduce-based distributed system which supports iterative data processing efficiently.

Keywords: Big data, MapReduce-based distributed systems, iterative data processing.

1 Introduction

Recently, research on big data has been actively made because big data are generated in various scientific applications, such as biology and astronomy. Because big data are a collection of data sets which are large and complex, it is difficult to process big data using on-hand database management tools or traditional data processing applications. Therefore, a lot of researches on a MapReduce framework have been done to perform the efficient analysis and mining on the big data. The MapReduce is a software platform which was developed by Google in 2004 to aim at processing big data in a distributed computing environment. The MapReduce processes big data by using Map and Reduce functions which are usually deployed in functional programming. It is considered as a main project by Amazon, Yahoo, and Google and is deployed for the cloud computing test bed of Yahoo [2]. The typical MapReduce-based distributed computing systems include Google MapReduce [1] and Hadoop MapReduce [2, 3].

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_45,

Meanwhile, research on big data analysis using a MapReduce framework has actively been done in various scientific applications, such as biology and astronomy. For example, genome data analysis requires iterative processing because it draws a result by doing operations iteratively. But, the existing MapReduce platform is appropriate for non-iterative data processing because it derive a result by doing one phase of Map and Reduce functions. As a result, the existing MapReduce is inefficient for iterative data processing because it perform Map and Reduce functions iteratively. To solve the problem, HaLoop [4] and Twister [5] were proposed as MapReduce-based distributed systems to support iterative data processing. To the best of our knowledge, there is little work on the performance analysis of MapReduce-based distributed systems supporting iterative data processing. Therefore, it is difficult to choose a proper MapReduce-based distributed system for iterative data processing applications.

In this paper, we first describe the existing MapReduce-based distributed systems which support iterative data processing. In addition, we do the performance analysis of the existing distributed systems in terms of execution time for various scientific applications which require iterative data processing. Finally, based on the performance analysis, we discuss some requirements for a new MapReduce-based distributed system which supports iterative data processing efficiently.

This paper is organized as follows. In Section2, we introduce typical MapReduce-based distributed data processing systems, such as Hadoop, HaLoop, and Twister. In Section 3, we do the performance analysis of the existing distributed systems for various scientific applications. In section 4, we discuss some requirements for a MapReduce-based distributed system. Finally, we draw our conclusion and suggest future work in Section 5.

2 Related Work

The representative distributed systems for iterative processing include Hadoop [2, 3], HaLoop [4], and Twister [5]. First, Hadoop developed by Apache group is the most popular system for processing big data and is provided in open source [2, 3]. It can deal with big data based on parallel processing by using a MapReduce framework. The MapReduce framework is a distributed programming technique proposed by Google for large-scale data processing in distributed computing environments. In MapReduce, the input computation is a list of (*key, value*) pairs and each map function produces intermediate (*key, value*) pairs. The framework groups the intermediate (*key, value*) pairs based on a hashing mechanism into the buckets of reduce tasks. The reduce tasks take both an intermediate key and a list of values as input and produce zero or more output results. Because Hadoop can support linear scalability and the fault tolerance of computing nodes, it is used as a platform by many companies and researchers for analyzing big data.

Second, HaLoop [4] is an enhanced version of the Hadoop framework, which was designed to support iterative processing. In addition, HaLoop dramatically improves their efficiency by making the task scheduler loop-aware and by adding various caching mechanisms. HaLoop not only handles loop control, but also offers a programming interface to express an iterative data analysis. The task scheduler of

HaLoop enables data reuse across iterations, by physically co-locating tasks that process the same data in different iterations. HaLoop caches and indexes data that are invariant across iterations in cluster nodes during the first iteration of an application. Caching the invariant data reduces the I/O cost for loading and shuffling them in subsequent iterations. HaLoop can avoid the need for a dedicated map-reduce step for fix-point or convergence checking by caching and indexing a reducer's local output.

Finally, Twister [5] is a distributed in-memory MapReduce runtime optimized for iterative MapReduce computations. It uses a publish/subscribe messaging infrastructure for communication and data transfers. Twister can support long running map/reduce tasks, which can be used in "configure once and use many times" approach. In addition, it manages invariant data for supporting efficient iterative MapReduce computations and provides programming extensions to MapReduce with the data transfers of broadcast and scatter types. These improvements allow Twister to support iterative MapReduce computations more efficiently compared to other MapReduce frameworks.

3 Performance Analysis

We compare the performance of the MapReduce-based distributed systems for iterative processing applications, namely Hadoop and HaLoop. HaLoop ver.1.2 and Hadoop ver.0.20.0 are used in our performance comparison because HaLoop is a modified version of Hadoop ver.0.20.0 to serve the iterative processing. By choosing these versions, we can fairly analyze the improvement of HaLoop with respect to the iterative processing. For performance analysis, we constitute a PC cluster which consists of 9 nodes. One of them is a name node to act as a master node and the other nodes are data nodes which act as slave nodes. Each node with Intel i5 2.8 Ghz CPU and 2GB of RAM operates on Ubuntu 12.04. We set the number of map and reduce tasks to five and the other system parameters to the default values.

We run three representative iterative applications, PageRank algorithm, descendant query, and k-mean algorithm, and compare the execution time on Hadoop and HaLoop. We use *Livejounal* data for the PageRank algorithm and descendant query, while the real map data of North-East America (NE) is used for the k-means algorithm. Table 1 shows the characteristics of the data sets used in our performance analysis.

Table 1. Data sets

name	*# of nodes*	*# of edges*	*size*
Livejournal	4.8 million	69 million	1GB
NE	0.12 million	-	2.9MB

3.1 Iterative Processing Applications

In this section, we describe iterative processing applications that are used in our performance analysis.

PageRank Algorithm. PageRank algorithm [6] is the core part of the Google search engine. The algorithm is being applied to the recommendation systems of the various fields. Given a graph consisting of nodes (e.g., web pages) and links among nodes, PageRank algorithm iteratively calculates a rank (weight) of each node based on the rank of its neighboring nodes. PageRank algorithm can be expressed into two steps, i.e., join step and merge step, on a MapReduce framework. In join step, a MapReduce job is required to join both node rank table and node linkage table. In merge step, another MapReduce job is defined to compute the rank of each node by considering the rank of its neighboring nodes. The algorithm terminates when the ranks have converged or the predefined number of iterations have been exceeded.

Descendant Query. Descendant query [4] increasingly retrieves the neighboring nodes from the given input node. The algorithm is being utilized in a clustering algorithm for astronomy and in a reachability algorithm for social network applications. Given a graph which consists of nodes (e.g., particles, friends) and link among nodes, descendant query iteratively finds all nodes that are within the predefined distance threshold (e.g., hop, distance) from an input node. Descendant query can be expressed into two steps, join step and duplicate elimination step. In join step, a MapReduce job is required to join query (*input*) node table and node linkage table. In duplicate elimination step, another MapReduce job is defined to delete the nodes found in this step which can be found in the input node table. The algorithm terminates when the number of retrieved nodes have converged or the predefined number of iterations have been exceeded.

***k*-means Algorithm.** *k*-means algorithm [4] is one of the typical clustering algorithms. The algorithm is being used in many scientific fields carrying out the data analysis tasks. Given a set of items, k-means algorithm performs clustering based on the mean value of items included in each generated cluster. *k*-means algorithm requires a MapReduce job to find the nearest cluster for each item and to assign the item to the cluster. The algorithm terminates when the mean value of each cluster does not differ from the value of the previous iteration or the predefined number of iterations have been exceeded.

3.2 Performance Comparison between Hadoop and HaLoop

Figure 1 shows the execution time of the PageRank algorithm with varying the number of iterations. HaLoop shows about 32% better performance than Hadoop. In the first iteration, the improvement is less than that of other iterations. The reason is that in the first iteration, HaLoop caches the sorted input data on each reducer's local disks and creates an index for the cached data. The main reason that HaLoop outperforms Hadoop is that HaLoop caches the invariant data being repeatedly used in all the iterations into reducer input cache. By doing so, the amount of transferred data in shuffling phase and the overall disk I/O are dramatically reduced, consequently leading to high performance improvement.

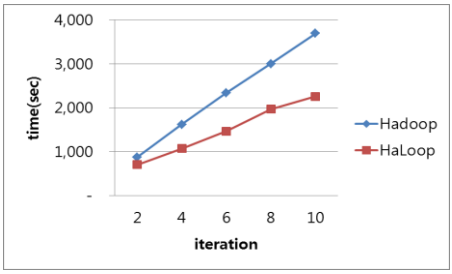


Fig. 1. The execution time of PageRank algorithm

Figure 2 shows the execution time of the descendant query with varying the number of iterations. HaLoop shows about 18% better performance than Hadoop. The improvement of HaLoop in descendant query is less than the improvement achieved in PageRank algorithm. The reason is that *Livejournal* data used in descendant query has a characteristic that a node in *Livejournal* data has many neighboring nodes. As a result, the descendant query in later iterations produces many duplicates which need to be eliminated. Consequently, the cost of HaLoop is dominated by the duplicate elimination step which does not gain much benefit from the HaLoop's caching mechanism. Thus, the improvement of HaLoop in descendant query is less than that achieved in PageRank algorithm.

Figure 3 shows the execution time of the k-means algorithm with varying the number of iterations. HaLoop shows about 19% better performance than Hadoop even though HaLoop cannot take advantage of reducer input cache for this application. The reason is that HaLoop can utilize the mapper input cache because the input data executed by the each mapper in the k-means algorithm is invariant in all the iterations. By caching the mapper input data, HaLoop can improve performance on overall execution time by reducing the amount of transferred data in a shuffling phase.

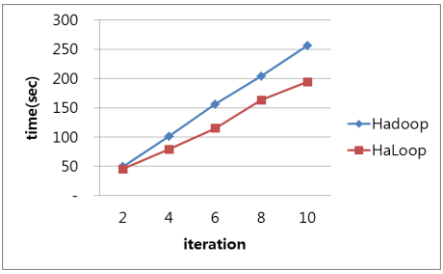


Fig. 2. The execution time of descendant query

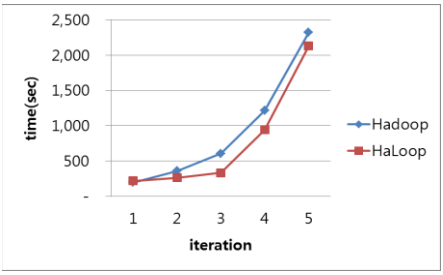


Fig. 3. The execution time of k-means algorithm

4 Discussion

HaLoop improves Hadoop in terms of the computational cost for iterative processing applications. This is because HaLoop can reduce data transmission cost and disk I/O cost by using both reducer input cache and mapper input cache. The cache mechanism of HaLoop efficiently deals with invariant data which are static in all the iterations, leading to better performance on execution time than Hadoop.

Through our performance experiment, we draw some requirements for improving the performance of HaLoop and Hadoop. First, the most of iterative processing applications requires the join computation, so it is necessary to handle input data with the mutually different forms. However, because both Hadoop and HaLoop receive an input data with one format, users who want to run the MapReduce job have to transform their input data. Thus, it is necessary to revise the existing MpaReduce-based distributed system so that they can receive input data with different formats. Second, because HaLoop and Hadoop do not have user friendly application program interface (API), user who want to run the MapReduce job have to fully understand both the MapReduce framework and their iterative applications. Therefore, it is necessary to provide a user-friendly API for the existing MpaReduce-based distributed systems so that scientists can execute their programs without the detailed knowledge of the MapReduce framework. Finally, the performances of the iterative processing applications are highly dependent on various system parameters in the existing MpaReduce-based distributed systems. However, users who want to run the MapReduce job scarcely know that are the important parameters for executing their iterative processing applications. Moreover, they do not know well how each parameter has to be set up for the optimized performance. Therefore, it is required to study on the optimization of system parameters for iterative applications.

5 Conclusion

In this paper, we analyzed the execution performances of the existing MapReduce-based distributed systems for various scientific applications which require iterative data processing. Through our performance analysis, HaLoop showed about 20% better performance than Hadoop. This is because HaLoop reduces data transmission and disk I/O costs by using both reducer input cache and mapper input cache. The cache mechanism of HaLoop can efficiently deals with invariant data which are static in all the iterations, consequently leading to better performance than Hadoop.

Finally, we present some requirements for a new MapReduce-based distributed system which supports iterative data processing efficiently. Based on the requirements, we will study on a scheme to receive heterogeneous input files. We will also develop user-friendly APIs which make it possible to easily write iterative programs for scientists who don't know much about the MapReduce platform.

Acknowledgement. This Research has been performed as a collaborative research project of project Building the System for Sharing and Convergence of Scientific Big Data and supported by the KOREA INSTITUTE of SCIENCE and TECHNOLOGY INFORMATION (KISTI).

References

1. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Operating System Design and Implementation, 10 (2004)
2. Apache Software Foundation, Apache Hadoop, http://hadoop.apache.org/
3. Apache Software Foundation, Hadoop Map- Redce, http://hadoop.apache.org/mapreduce
4. Bu, Y., Howe, B., Balazinska, M., Ernst, M.D.: HaLoop: Efficient Iterative Data Processing on Large Clusters. In: VLDB (2010)
5. Ekanayake, J., Li, H., Zhang, B., Gunarathne, T., Bae, S.- H., Qiu, J., Fox, G.C.: Twister: A Runtime for Iterative MapReduce. In: The ACM International Symposium on High Performance Distributed Computing, HPDC (2010)
6. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab (1999)

A Semi-clustering Scheme for Large-Scale Graph Analysis on Hadoop

Seungtae Hong[1], Youngsung Shin[1], Dong Hoon Choi[2], Heeseung Jo[1], and Jae-woo Chang[1,*]

[1] Dept. of Computer Engineering, Chonbuk National University, Jeonju, South Korea
{dantehst,twotoma,heeseung,jwchang}@jbnu.ac.kr
[2] Korea Institute of Science and Technology Information (KISTI), Daejeon, South Korea
choid@kisti.re.kr

Abstract. With the evolution of IT technologies, large-scale graph data have lately become a growing interest. As a result, there are a lot of research results in large-scale graph analysis on Hadoop. The graph analysis based on Hadoop provides parallel programming models with data partitioning and contains iterative phases of MapReduce jobs. Therefore, the effectiveness of data partitioning depends on how the data partitioning maintains data locality in each node of cluster. In this paper, we propose a semi-clustering scheme for large-scale graph analysis such as PageRank algorithm on Hadoop and show that the proposed scheme is effective. With experiment results, PageRank computation with the semi-clustering improves the performance.

Keywords: large-scale graph analysis, semi-clustering, Hadoop, PageRank.

1 Introduction

With the evolution of IT technologies, large-scale graph data, such as web graph, social network, bio network, have lately become a growing interest in the commercial world and research communities. Therefore, researches on parallel programming models for effective analysis and mining on large-scale graph data have been actively studied. There are various kinds of parallel programming models that can be used for analyzing the large-scale graphs, for example Hadoop [1], Pregel [2], and M3R [3]. Among them, Hadoop is widely adopted due to its easiness of programming through simple API with Map and Reduce. As a result, there are a lot of research results in large-scale graph analysis on Hadoop.

The graph analysis based on Hadoop provides parallel programming models with data partitioning and contains iterative phases of MapReduce jobs. Each iterative phase creates a new MapReduce job, and the previous MapReduce job conveys its results data into the input data of the next MapReduce job. At this stage, a MapReduce job shuffles the data of each node through network communication among nodes. As much of previous research pointed out that the network communication overhead was largely dependent on the overall performance of MapReduce processing [3, 4], it is essential to

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_46,

minimize the exchange cost among nodes by optimizing the data partitioning for the performance of graph analysis based on Hadoop. The effectiveness of data partitioning depends on how the data partitioning maintains data locality in each node of cluster.

In this paper, we propose a semi-clustering scheme for large-scale graph analysis on Hadoop and show that the proposed scheme can be effective such as PageRank algorithm. The rest of the paper is organized as follows. Section 2 presents related work and Section 3 describes the details of our semi-clustering scheme and brief algorithms. In section 4, we demonstrate the performance analysis of our semi-clustering scheme through referenced data and PageRank algorithm. Finally, Section 5 concludes our work and presents a future direction.

2 Related Work

The typical graph analysis techniques which calculate the importance of each node in the graph data are as follows. The PageRank [5] algorithm, which is proposed for link analysis of web graphs, is commonly used for clustering [6] and social network analysis [7]. The algorithm can be useful when calculating the importance of each node or the impact to the other parties in social network, and also used for calculating the value of each node in protein interaction network analysis in bioinformatics [8]. The HITS algorithm is widely used in calculating the ranks of each node of hypertext [9]. However, The HITS algorithm is known to show lower sensibility than the PageRank algorithm [10]. For example, a rank value is hard to be altered even when a structure of graph is changed. In bioinformatics, the PageRank algorithm is more useful to evaluate the importance of each protein, since false positive and false negative edges occur frequently in protein interaction networks.

Hadoop provides two partitioning technique for data parallelism; hash-based and range-based. The hash partitioning technique splits the graph data based on fixed hash values. Therefore, it is inefficient because it does not consider the connectivity of graph data. On the other hand, although the range partitioning technique can maintain the connectivity of graph data, it is limited to specific issues.

Lin [11] partitioned input data based on the range partitioning technique grouping related vertices to a data partition according to domain names. Although the scheme shows improved performance of PageRank, the range partitioning scheme based on domains is hard to be applied in web graph which comes from a single domain.

3 Semi-clustering Scheme

In this section, we describe the detailed algorithms of the proposed semi-clustering scheme for large-scale graph analysis on Hadoop. In the proposed scheme, for each vertex which contains a lot of outgoing edges, the semi-clustering creates a group of reachable vertices that starts traversing from it. By allocating a group of reachable vertices into the same data partition, it reduces the cost of shuffling and performs efficient computation in graph analysis on the Hadoop.

3.1 Sort of Vertices in a Graph

Map phase calculates the number of outgoing edges of each vertex and outputs data of the form <the number of outgoing edge, vertex ID> into local disks. The outputs from

Map function are sorted in decreasing order by the number of outgoing edges since MapReduce framework automatically sorts output data by keys on shuffling. Reduce function outputs the sorted list into HDFS. A group of reachable vertices starting from the vertex at the beginning of the output creates a set of semi-clusters for a given graph. Figure 1 shows the sorting vertices algorithm.

```
1:  Class Mapper  // computing the number of outgoing edges of vertices //
2:    Method Map(nid n, node N)
3:      count <- 0
3:      for all nodeid m∈N.AdjacencyList do
4:        count++
5:      emit(count, nid n)

1:  Class Reducer  // sorting //
2:    Method Reducer(outdegree count, nids [n1,n2, …])
3:      for all nid ni ∈ nids [n1,n2, …] do
4:        emit(outdegree count, nid ni)
```

Fig. 1. Sorting vertices algorithm

3.2 Creation of Semi-clusters

Starting from the vertex which has the largest number of outgoing edges in the sorted list, the graph is traversed in breadth-first search way to insert the neighboring vertices into a set of reachable vertices. This process is repeated until no newly reachable vertices are left. Each set of reachable set is called a semi-cluster, and the number of

```
1:  Class SemiClusterDriver
2:    Method Main(depth k)
3:      numIteration <- 0
4:      for numIteration <k do
5:        Map()
6:        Reduce()
7:        numIteration++

1:  Class Mapper
2:    Method Map(nid n, node N)
3:      c <- N.ClusterID
4:      if c == null then
5:        c <- findGroup(n)
6:      for all nodeid m∈N.AdjacencyList do
7:        if c != null  then
8:          emit(nid m,c)
9:      emit(nid n, N)

1:  Class Reducer
2:    Method Reducer(nid m, [c1,c2,…])
3:      M <-  , p<-0
4:      for all c ∈ [c1,c2, …] do
5:        if IsNode(c) then
6:          M <- c
7:        else if p == 0 then
8:          p <- c
9:      M.ClusterID <- p
10:     emit(nid m, node M)
```

Fig. 2. Creating semi-clusters algorithm

vertices that organizes a semi-cluster is called cardinality. Cardinality is utilized to limit the size of a partition when allocating semi-clusters to partitions. Figure 2 shows the creating semi-clusters algorithm.

3.3 Allocation of Semi-clusters to Partitions

Starting from the semi-clusters with high cardinality, semi-clusters are allocated to data partitions in round-robin way. If the cardinality of a semi-cluster exceeds the size of data partition, the semi-cluster is allocated to the other partition that has enough capacity. Figure 3 shows the allocating semi-clusters algorithm.

```
1:  Class Mapper
2:    Method Map(nid n, node N)
3:      M <- n+N
4:      emit(N.ClusterID, M)

1:  Class Partitioner
2:    Method getPartition(clusterID c, node M)
3:      p<-getPartitionID(c)
4:      for IsSuitable(p) != true do
5:        p<-getPartitionID(c)
6:      return p
```

Fig. 3. Allocating semi-clusters algorithm

4 Performance Analysis

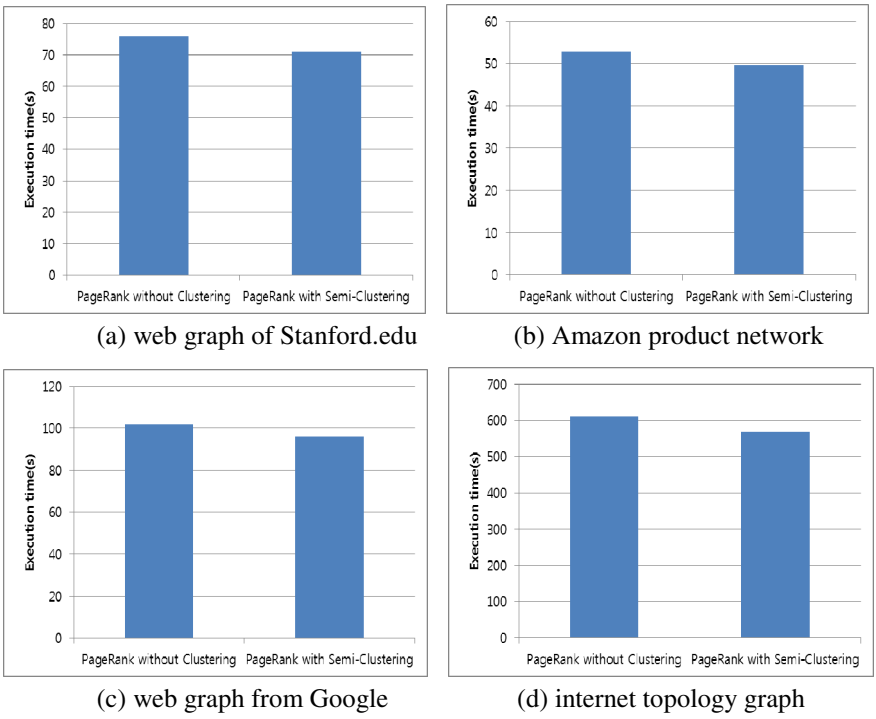
In this section, we evaluate the proposed semi-clustering scheme with PageRank algorithm. Table 1 denotes the evaluation environment for PageRank based on semi-clustering. Joycrawler 0.2 [12] is used as a PageRank algorithm implementation in this evaluation. We evaluate our semi-clustering scheme by using four graph datasets: web graph of Stanford.edu[13] (281,903 nodes, 2,312,497 edges), Amazon product network[14] (334,863 nodes, 925,872 edges), web graph from Google[13] (875,713 nodes, 5,105,039 edges), and internet topology graph[15] (1,696,415 nodes, 11,095,298 edges).

Table 1. Physical cluster

Number of physical machines	9 computing servers(Master : 1, Slave : 8)
CPU	Intel i5.3.4GHz
Main Memory	2GB
HDD	500GB
OS	Ubuntu 12.04

Fig. 4 shows the evaluation results of PageRank algorithm with four datasets. We measure the execution time of both the PageRank algorithm without semi-clustering and with semi-clustering. In the case of the internet topology graph, the execution

time of PageRank algorithm with semi-clustering is reduced by 7.2%. This evaluation results denote PageRank computation with the semi-clustering scheme improves the performance, as the number of nodes and edges increase. Therefore, we can confirm that the proposed scheme is effective for large-scale graph processing.



(a) web graph of Stanford.edu (b) Amazon product network

(c) web graph from Google (d) internet topology graph

Fig. 4. PageRank execution time

5 Conclusion and Future Work

In this paper, we propose a semi-clustering scheme for large-scale graph analysis on Hadoop. In our proposed scheme, for each vertex which contains a lot of outgoing edges, the semi-clustering creates a group of reachable vertices that starts traversing from it. By allocating a group of reachable vertices into the same data partition, the proposed scheme reduces the cost of shuffling and performs efficient computation in graph analysis on the Hadoop. With experiment results, PageRank computation with the semi-clustering improved the performance, as the number of iterations increases. Therefore, our scheme can be largely effective for large-scale graph analysis. As a future work, we plan to evaluate our scheme on much more number of cluster nodes.

Acknowledgement. This Research has been performed as a collaborative research project of project Building the System for Sharing and Convergence of Scientific Big Data and supported by the KOREA INSTITUTE of SCIENCE and TECHNOLOGY INFORMATION (KISTI).

References

1. Hadoop, http://hadoop.apache.org/
2. Malewicz, G., Austern, M., Bik, A., Dehnert, J., Horn, I.: Pregel: a system for large-scale graph processing. In: SIGMOD 2010 (2010)
3. Shinnar, A., Cunningham, D., Herta, B., Saraswat, V.: M3R: Increased performance for in-memory Hadoop jobs. In: VLDB 2012 (2012)
4. Bu, Y., Howe, B., Balazinska, M., Ernst, M.D.: HaLoop: Efficient iterative data processing on large clusters. In: VLDB 2010 (2010)
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: WWW 1998 (1998)
6. Avrachenkov, K., Dobrynin, V., Nemirovsky, D., Pham, S., Smirnova, E.: PageRank based clustering of hypertext document collections. In: SIGIR 2008 (2008)
7. White, S., Smyth, P.: Algorithms for estimating relative importance in networks. In: KDD 2003 (2003)
8. Ivn, G., Grolmusz, V.: When the web meets the cell: Using personalized PageRank for analyzing protein interaction networks. Bioinformatics Advance Access (December 2010)
9. Kleinberg, J.: Authoritative sources in a hyperlinked environment. JACM 46(5), 604–632 (1999)
10. Lee, H.C., Borodin, A.: Perturbation of the hyperlinked environment. In: Warnow, T.J., Zhu, B. (eds.) COCOON 2003. LNCS, vol. 2697, pp. 272–283. Springer, Heidelberg (2003)
11. Lin, J., Schatz, M.: Design pattern for efficient graph algorithms in MapReduce. In: MLG 2010 (2010)
12. Joycrawler, http://code.google.com/p/joycrawler/
13. Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M.: Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics (2009)
14. Yang, J., Leskovec, J.: Defining and Evaluating Network Communities based on Ground-truth. In: ICDM (2012)
15. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD (2005)

Multi-stream Parallel String Matching on Kepler Architecture

Nhat-Phuong Tran[1], Myungho Lee[1,*], Sugwon Hong[1], and Dong Hoon Choi[2]

[1] Department of Computer Science and Engineering, Myongji University,
38-2 San Namdong, Cheo-In GuYong In, Kyung Ki Do, Korea 449-728
[2] Korea Institute of Science and Technology Information (KISTI),
245 Dae Hak Ro, Yu Seong Gu, Daejeon, Korea 305-806

Abstract. Aho-Corasick (AC) algorithm is a commonly used string matching algorithm. It performs multiple patterns matching for computer and network security, bioinformatics, among many other applications. These applications impose high computational requirements, thus efficient parallelization of the AC algorithm is crucial. In this paper, we present a multi-stream based parallelization approach for the string matching using the AC algorithm on the latest Nvidia Kepler architecture. Our approach efficiently utilizes the HyperQ feature of the Kepler GPU so that multiple streams generated from a number of OpenMP threads running on the host multicore processor can be efficiently executed on a large number of fine-grain processing cores. Experimental results show that our approach delivers up to 420Gbps throughput performance on Nvidia Tesla K20 GPU.

Keywords: string matching, Kepler GPU, multi-stream, HyperQ, multithreading.

1 Introduction

Aho-Corasick (AC) algorithm [1] is a multiple patterns matching algorithm which can simultaneously match a number of patterns for a given finite set of strings (or dictionary) against a given input data. The AC algorithm is commonly used in various applications such as network intrusion detection [16], [17], genome/protein matching for bio-sequence analysis [12], [15], among many others. In order to speed up the string matching operations and meet the real-time performance requirement imposed on these applications, achieving high performance for the AC algorithm is crucial.

Recently, the Graphic Processing Unit (GPU) is becoming increasingly popular for various applications. The architecture of the GPU has gone through a number of innovative design changes in the last decade which have drastically increased the peak floating-point throughput performance (flops) [5], [6]. Although the AC algorithm is not floating-point intensive, it can be benefitted by many promising architectural characteristics of the GPU. A GPU has a large number of (fine-grain) cores where massive parallel pattern matching operations for the AC algorithm can be performed in parallel. A GPU also provides high memory bandwidths. Thus it can

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_47,

feed the input data and the reference pattern data at a high rate for possible matches. On the other hand, a GPU has a complicated memory hierarchy whose efficient use has a major effect on the application's performance and is mostly under the programmer's control. Therefore, we need sophisticated parallelization techniques to achieve high performance for the AC algorithm.

In this paper, we present a multi-stream based parallelization approach for the AC algorithm on the latest Nvidia Kepler GPU. Our approach efficiently utilizes the HyperQ feature of the Kepler GPU so that multiple streams generated from a number of OpenMP threads running on the host multicore processor can be efficiently distributed and executed on a large number of fine-grain processing cores. furthermore, it also exploits the high degree of the on-chip parallelism and the complicated memory hierarchy of the Kepler GPU in order to maximize the throughput performance. Experimental results on Nvidia Tesla K20 GPU based on Kepler GK110 architecture along with multicore host processor (Intel Xeon E5-2650) show that our approach delivers up to 420Gbps throughput. Comparing with a single stream parallelization approach, it leads to 1.45-times higher throughput performance.

The rest of the paper is organized as follows: Sections 2 introduces the AC algorithm. Section 3 describes the architecture of the latest GPU including the Nvidia Kepler and its execution model. Section 4 explains our multi-stream based parallelization approach of the AC algorithm on the Kepler architecture. Section 5 shows the experimental results on Nvidia Tesla K20 GPU employing the Kepler GK110 architecture. Section 6 wraps up the paper with conclusions.

2 Aho-Corasick (AC) Algorithm

The Aho-Corasick (AC) algorithm is a multiple patterns matching algorithm which can match multiple patterns simultaneously for a given finite set of strings (or dictionary). The AC algorithm can be implemented as a Non-deterministic Finite Automata (NFA) or a Deterministic Finite Automata (DFA). The AC consists of two phases: 1) first, a pattern matching machine called the AC automaton (machine) is constructed from a finite set of patterns; 2) second, the input text data is applied to the constructed AC machine in order to find the locations that the patterns appear [1].

The AC automaton invokes three functions: a goto function g, a failure function f, and an output function *output*:

- The goto function g function maps a pair consisting of a state and an input symbol into a state or a message *fail*. The AC machine has the property that $g(0, \sigma) \neq fail$ for all input symbol σ.
- The failure function f maps a state into another state. It is consulted whenever the goto function reports a "fail".
- The output function *output* maps a set of keywords to output at the designated states.

We implement the AC algorithm as a DFA. The DFA consists of a finite set of states S and a next move function δ such that for each state s and an input symbol a, $\delta(s,a)$ is a state in S [4]. Thus, the next move function δ is used in place of both the goto function and the failure function. The output function is also incorporated in the DFA. The DFA processes the input text with length n in $O(n)$.

3 Overview of Nvidia Kepler GPU Architecture

In this paper, we use Nvidia's latest GPU based on Kepler architecture (Tesla K20) consisting of 15 Streaming Multiprocessors (SMXs) or thread blocks with the Compute Capability 3.5. Compared with the previous Fermi architecture, it has a larger number of threads and registers available on each SMX. Furthermore, it provides new architectural features such as Hyper-Q, Dynamic Parallelism, and GPUDirect [19]. The Dynamic Parallelism adds the capability for the GPU to generate new work for itself, synchronize the results, and control the scheduling of the work without the involvement of the CPU. Thus it provides the flexibility to adapt to the amount and the form of parallelism through the program execution. The GPUDirect enables GPUs within and outside a single computer to directly exchange data without going through the CPU and system memory. Thus can significantly reduce the data transfer overheads. The Hyper-Q allows multiple CPU cores to launch work on a single Kepler GPU simultaneously. This increases the utilization of the GPU, thus improve the throughput performance with the involvement of multiple CPU cores. In this paper, we utilize the Hyper-Q feature to maximize the throughput performance of the AC algorithm.

For executing programs on the Nvidia GPU, we use CUDA. CUDA programs use a hierarchy of memories of the Nvidia's GPU. They are registers and local memories belonging to each thread, a shared memory and the level-1 data cache used in a thread block (SMX in Kepler architecture) and shared by threads belonging to the block, and the global memory accessed from all the thread blocks [5, 6]. In CUDA programs, data needed for computations on the GPU is transferred from the host memory to the global memory, optionally placed in the shared memory by the programmer or automatically loaded in the L1 cache or read-only data cache by the cache controller, and used by thread blocks and thread processors through the registers. The multiple threads assigned to each thread block executes in the SIMD mode by having the same instruction managed by the Instruction Unit on different portions of data. When a running thread encounters a cache miss, for example, the context is switched to a new thread while the cache miss is serviced for the next 400 cycles or more. Thus the GPU is executing in a multithreaded fashion.

4 Multi-stream Parallelization Approach

The AC algorithm introduced in Section 2 proceeds in two steps: 1) construction of the AC pattern matching machine; 2) conducting the pattern matching operations using the machine. In typical pattern matching applications using the AC, the first phase is performed once and the second phase is repeated multiple times. In this paper, we perform the first phase of the AC sequentially using single CPU core. Then we perform the second phase on the GPU in parallel with the involvement of multiple CPU cores where multiple streams are generated. Thus the multi-stream parallelization approach is focused on the second phase.

In AC algorithm, the input text length is usually very long. Thus we partition the input text into many parts and apply the multiple patterns matching procedures in each part in parallel. In order to process the pattern matching in each part, we generate a CUDA stream. Thus multiple CUDA streams are mapped onto single Kepler GPU using the new Hyper-Q feature. In the earlier Fermi GPU, up to 16 streams are mapped to a GPU. However, all the streams are multiplexed into the same hardware work queue. Thus they are executed serially in the same queue. In the Kepler architecture, there are up to 32 hardware work queues between the host and the CUDA Work Distributor (CWD) logic in the GPU. Thus we can generate up to 32 streams and map the streams onto the same GPU to run them concurrently (see Figure 1).

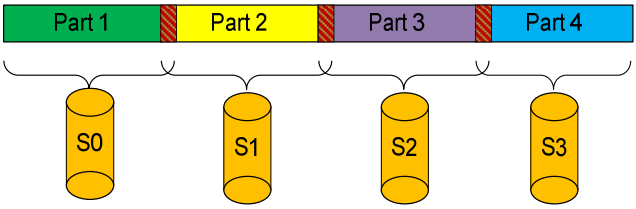


Fig. 1. Partitioning of input data into multiple parts to create multiple CUDA streams for parallel execution on the Kepler GPU using the Hyper-Q

Each stream calls the kernel function independently for matching patterns on each part of input data corresponding to each stream. Each kernel function creates a number of blocks and a number of threads per block for applying the pattern matching. Compared with the Fermi architecture where the multiple streams are time-multiplexed to share the same GPU, the Kepler architecture allows the sharing of the GPU simultaneously [19]. Thus, resources of the Kepler GPU are utilized more efficiently. In order to implement the parallel multi-stream pattern matching, we create a number of OpenMP threads on the host multicore processor each of which create a stream individually. Each thread copy parts of the input data asynchronously to the global memory while the pattern matching is performed on the GPU. Thus, the kernel execution and the data transfer can be overlapped (see Figure 2). This improves the application's performance. The pinned memory on the host memory is used for the asynchronous copy. Thus, the data in host memory must be page-locked memory.

```
#pragma omp parallel for …
for(int i = 0; i < N ; i++)
{
    cudaMemcpyAsync( dev1, host1, size1,
                     cudaMemcpyHostToDevice, stream[i] );
    kernel<<<gdim, bdim, smem, stream[i]>>>( <parameters> );
    cudaMemcpyAsync( host2, dev2, size2,
                     cudaMemcpyDeviceToHost, stream[i] );
}
```

Fig. 2. Code snippets for creating and running parallel multi-streams

5 Experimental Results

We implemented the parallel multi-stream AC algorithm. Our experiments are conducted on a system including the Intel multicore processor (2.0Ghz Intel Xeon E5-2650) with 20MB level-3 cache, Nvidia Tesla K20 GPU with 5GB device memory. We also used Nvidia GeForce GTX 285 GPU for performance comparison with the K20 GPU. The OS is Centos 5.5. We used input data sizes in the range of 50KB - 500MB and the numbers of patterns in the range of 100 – 20,000. In order to generate the random input data sets and the reference pattern data sets, we first collected 50GB of data from a variety of magazines such as TIME, BBC, among many others. Then we extracted the input data and the pattern data from the collected data. In all experiments conducted, we ignored the time spent in the construction phase of STT which run on single CPU core and the time to copy the input text data and the STT to the GPU device memory.

Figure 3 show the throughput performance of different input sizes when the number of patterns is fixed at 20000. The throughput increases as the data size increases, in general. This is especially true for the Kepler GPU where the number of cores has drastically increased compared with the previous generation GPUs. Our approach delivers up to 420Gbps throughput. Comparing with a single stream parallelization approach, it leads to 1.45-times higher throughput performance. The throughput on the GTX285, however, is rather flat as the data size increases: the throughput saturates around 100Gbps.

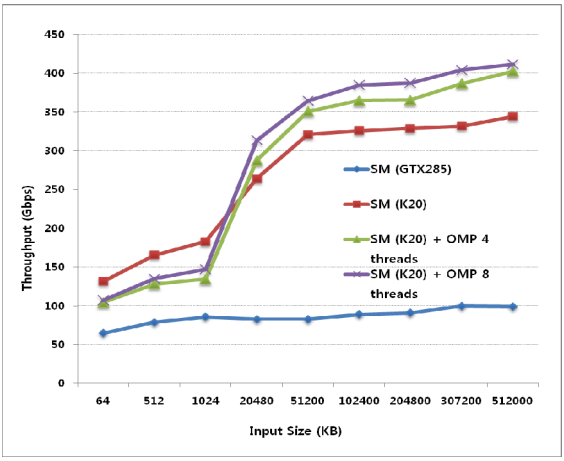


Fig. 3. Throughput (Gbps) for different input data sizes when the number of patterns is fixed at 20000

6 Conclusions

In this paper, we proposed a multi-stream parallelization approach for the AC algorithm on a GPU. The proposed approach efficiently utilizes the HyperQ feature of the Kepler GPU so that multiple streams generated from a number of OpenMP

threads running on the host multicore processor can be efficiently distributed and executed on a large number of fine-grain processing cores. Experimental results on Nvidia Tesla K20 GPU based on Kepler GK110 architecture along with multicore host processor (Intel Xeon E5-2650) show that our approach delivers up to 420Gbps throughput. Comparing with a single stream parallelization approach, it leads to 1.45-times higher throughput performance.

Acknowledgements. The authors would like extend their thanks to the Center for Computing and This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (Grant No: 2012-042269).

References

1. Aho, A.V., Corasick, M.J.: Efficient string matching: An aid to bibliographic search. Communications of the ACM 20(Session 10), 761–772 (1977)
2. Jacob, N., Brodley, C.: Offloading IDS Computation to the GPU. In: The 22nd Annual Computer Security Applications Conference (2006)
3. Lin, C.-H., Tsai, S.-Y., Liu, C.-H., Chang, S.-C., Shyu, J.-M.: Accelerating String Matching Using Multi-Threaded Algorithm on GPU. In: 2010 IEEE Global Telecommunications Conference, GLOBECOM 2010, December 6-10, pp. 1–5 (2010)
4. Norton, M.: Optimizing Pattern Matching for Intrusion Detection (July 2004), http://docs.idsresearch.org/OptimizingPatternMatchingForIDS.pdf
5. NVIDIA, CUDA Best Practices Guide: NVIDIA CUDA C Programming Best Practices Guide – CUDA Toolkit 4.0 (May 2011)
6. NVIDIA, NVidia gtx280, http://kr.nvidia.com/object/geforce_family_kr.html
7. OpenACC (March 2012), http://www.openacc-standard.org
8. OpenCL, http://www.khronos.org/opencl/
9. Saavedra-Barrera, R.H., Culler, D.E., von Eicken, T.: Analysis of multithreaded architectures for parallel computing. In: ACM Symposium on Parallel Algorithms and Architectures - SPAA, pp. 169–178 (1990)
10. Scarpazza, D., Villa, O., Petrini, F.: Peak-Performance DFA-based String Matching on the Cell Processor. In: International Workshop on System Management Techniques, Processes, and Services (2007)
11. Scarpazza, D., Villa, O., Petrini, F.: Accelerating Real-Time String Searching with Multicore Processors. IEEE Computer Society (2008)
12. Schatz, M.C., Trapnell, C.: Fast Exact String Matching on the GPU. Center for Bioinformatics and Computational Biology (2007)
13. Sen, S.: Performance Charaterization and Improvement of Snort as an IDS (August 2006), http://www.princeton.edu/~soumyas/bell_labs_report_snort.pdf
14. Smith, R., Goyal, N., Ormont, J., Sankaralingam, K., Estan, C.: Evaluating GPUs for Network Packet Signature Matching. In: IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS 2009, April 26-28, pp. 175–184 (2009)
15. Tumeo, A., Villa, O.: Accelerating DNA analysis applications on GPU clusters. In: 2010 IEEE 8th Symposium on Application Specific Processors (SASP), June 13-14, pp. 71–76 (2010)

16. Tumeo, A., Villa, O.: Efficient Pattern Matching on GPUs for Intrusion Detection Systems. In: Proceedings of the 7th ACM International Conference on Computing Frontiers (2010)
17. Vasiliadis, G., Antonatos, S., Polychronakis, M., Markatos, E.P., Ioannidis, S.: Gnort: High Performance Network Intrusion Detection Using Graphics Processors. In: Lippmann, R., Kirda, E., Trachtenberg, A. (eds.) RAID 2008. LNCS, vol. 5230, pp. 116–134. Springer, Heidelberg (2008)
18. Volkov, V., Demmel, J.W.: Benchmarking GPUs to Tune Dense Linear Algebra. In: SC 2008, pp. Art.31:1–31:11 (November 2008)
19. White paper, NVIDIA's Next Generation CUDA Compute Architecture: Kepler GK 110 The Fastest, Most Efficient HPC Architecture Ever Built, Nvidia (2012)
20. Zha, X., Sahni, S.: Multipattern string matching on a GPU. In: IEEE Symposium on Computers and Communications (ISCC), June 28-July 1, pp. 277–282 (2011)
21. Zha, X., Scarpazza, D., Sahni, S.: Highly Compressed Multi-pattern String Matching on the Cell Broadband Engine. In: IEEE Symposium on Computers and Communications (ISCC), June 28-July 1, pp. 257–264 (2011)

Microscopic Bit-Level Wear-Leveling for NAND Flash Memory

Yong Song[1], Woomin Hwang[1], Ki-Woong Park[2], and Kyu Ho Park[1]

[1] CORE Lab., Dept. of Electrical Engineering, KAIST, Korea
[2] Dept. of Computer Hacking and Information Security, Daejeon University, Korea
{ysong,wmhwang,kpark}@core.kaist.ac.kr, woongbak@dju.kr

Abstract. By microscopically observing widely used data files, we identified the considerable room for life time improvement in NAND flash memory, which is due to the discovery of a non-uniformity in bit-level data patterns. In an attempt to exploit the discovery, we propose a novel bit-level wear-leveling scheme. Instead of considering only the view of page-level or block-level, we incorporate the non-uniformity in data encoding patterns into wear-leveling scheme. Because of its orthogonality to the existing block-level wear-leveling approaches, our solution can be adopted over the existing solutions without considerable overhead and extend NAND flash's life span up to 36% in case of SLC.

1 Introduction

During the last decade, NAND flash memory has become a de facto solid state storage technology. In the NAND flash memory-based storage systems, it is essential to reliably store data for years, however, NAND flash has endurance problem that each NAND flash cell is worn-out by program/erase cycling and eventually loses its capability to store data. Thus, the endurance problem of NAND flash has been considered in many previous researches and they proposed wear-leveling algorithms, which even out the wearing of different blocks of NAND flash memory.

To enlarge NAND flash's life span, many researches proposed their algorithms for wear-leveling which tries to make wear-out of different blocks of flash memory even [3][4][5][6][7]. To expand lifetime of NAND flash blocks, commonly, the previous algorithms take page-granularity approaches and use block erase count as a wear-out indicator.This is based on an implcit assumption all cells in same page or same block have the same probability to be worn-out by P/E cycle. To identify whether the assumption is right or not, we obtained data patterns of widely used file data frequently written to disk or accounting for large part of storage, for example, linux source codes, database workloads, encoded video files. After analyzing the data, we found out a case that the bit-level data pattern is not uniform. Base on our observation, we build a simple statistical wear-out model of NAND flash cell and we propose bit-level wear-leveling scheme. The evaluation shows that it expands life span of NAND flash up to 30%.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing,* Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_48,

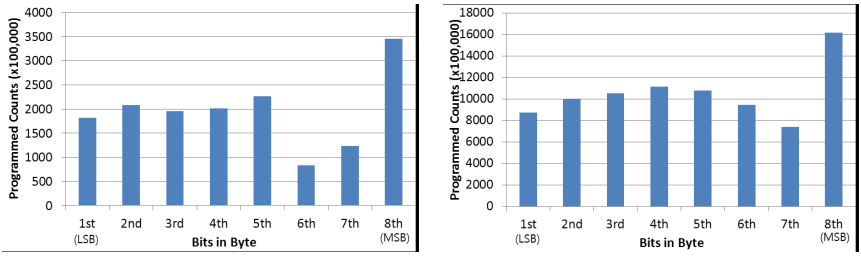


Fig. 1. (a) Bitwise counts of programmed state of Linux source codes, (b)Bitwise count of programmed state of DB workloads

2 Motivation

In case of SLC NAND flash, a cell stores a bit and a page is composed of a series of NAND cells and a block contains multiple pages. To store data on them, a block erase operation should be performed prior to write data, which makes all cells in the block changed to the erased state, representing logic '1'. If logical value of '0' is written to a cell when page writing operation of NAND flash, a state transition of the cell occurs from the erase state to the programmed state, indicating '0'. On the other hands, if logical value of '1' is written to a cell, there is no state transition of the cell and it remains in the erased state. Because a state transition from erase state to programmed state makes a cell worn out[2], each cell can experience different wear-out progress accoriding to the number of its actual programs.

A problem arises from that the previous block-based wear-leveling algorithms treat all cells comprising a block together as if they experience the same number of state transitions between the erase state and the programmed state, thus having the same degree of wear-out. If we assume that the physical endurance of each cell is same, among all cells comprising a block, the cell experiencing the highest number of programs is to be completely worn out most rapidly. If the number of completely worn-out cells is bigger than the number of correctable bit errors by bit error correction algorithm such as ECC or DHC, it is regarded that the life time of the page is ended. Even though other cells still have their life time left and they can operate normally, the page is not used any more. This leads to the underestimation of the life time of the target page. It therefore makes the NAND flash memory block lose chances to prolong lifetime of the block.

Outstanding freqeuncy of programs in a specific bit position appears significantly in text-based workloads. We selected Linux source codes as a representative text-based workloads and obtained their bit-level data patterns by traversing Linux source code tree. Figure 1(a) shows the accumulated counts of the case of logical '0' at each bit position per byte, which requires the state transition of the cell from erase state to the programmed state. We can see that the most significant bit (MSB) position in a byte(character) experiences the highest number

of the state transtion than other bit positions. This is because ASCII code for human readable data does not use MSB.

In order to observe data patterns of files to be freqently written to disk, we performed DBT2[1] benchmarks for SQLite[8], an open-source database. The test creates database workloads and simulate heavy user loads for OLTP. We log all of the contents synchronously written to disk through the test. We can see that the bit-level data pattern of the result obtained from the benchmark test is similar to the previous one, as shown in Figure 1(b). Consequently, the difference in degree of wear-out of each cell would be apart more as the ratio of ASCII characters increases and it provides us rooms for prolonging lifetime of the target NAND block.

3 Proposed Scheme

In this section, we present a bit-level wear-leveling scheme based on a simplified statistical model to span the life time of NAND flash memory, which is derived from the observation of the data non-uniformity as described in Section 2. First, we define the fundamental wear-out model of SLC cell, in accordance with the writing pattern of the data. Then, we present the proposed bit-level wear-leveling schemes to make the degree of wear-out of cells uniform.

Our Assumption: Our model is basically derived from the concept of the charge-to-breakdown, Q_{BD}[2]. The Q_{BD} is defined to amounts of charges crossing the tunnel oxide in a NAND flash cell until the oxide is broken down. It is from the following physical characteristics: the oxide of each NAND flash cell is little by little worn-out by repeated program/erase cycling, and the cell eventually loses its capability to store data. Base on the above, the following assumption is obtained. The degree of wear-out of a cell after a P/E cycling is proportional to the amount of charge passing the oxide in the cell and the number of injected and ejected charges during a cycling is always same, and threshold voltage for sensing stored data is also propotional to the amount of stored charges. This assumption is a base of the wear-out modeling for a SLC NAND cell.

SLC can store only one bit value per cell, which basically is a threashold voltage level. Typically, in NAND flash, a logical bit value '1' of an SLC cell denotes the erased state, and a logical bit value '0' denotes the programmed state. W_{slc} is defined to the degree of worn-out of the oxide by charge injection/ejection during a P/E cycling. In case of SLC, a cell having a bit value of '0(Programmed)' is worn out by W_{slc} in a P/E cycling. And a cell having a bit value of '1(Erased)' is not worn-out because there is no charge tunneling the oxide during the cycling. We simply assume that the degree of wear-out of a cell not in programmed state is zero because, in the case, there's no charge to be ejected from FG to substrate through the oxide. Based on the above modeling principles for SLC, the degree of wear-out of each cell is simply represented to the counts of the case that it has a bit value of 0.

In order to even out the usage of each cell, the technique for distributing the erasures and re-writes evenly across the cells is necessary. To achieve it, we

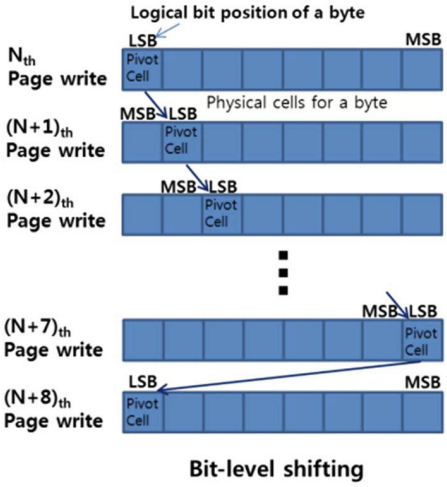


Fig. 2. Our proposed schemes for expanding lifetime of NAND flash cells

propose a simple bit-level shifting scheme for SLC as shown in Figure 2. To make the frequncy of programmed state of each cell statistically even, every page-level program and erase cycling, a byte of data is written from the pivot bit in the byte and the rest of bits are written from the LSB bit position. The location of pivot bit is shifted by one bit. In practical implementation of this scheme, we can set the location of pivot bit as the value of block erase counts modulo 8.

4 Evaluation

We evaluated our bit-level wear leveling scheme based on the proposed simple wear-out model for SLC NAND flash. The evaluation is based on the previously obtained data patterns for different file data workloads, Linux source code files, which are the most good reference for docuement file data composed of mostly ASCII code-based data, and DB update data, which is frequently update to the storage and have big influence on NAND flash's wear-out.

To simulate and visualize the page write operations onto NAND flash memory, we examine data pattern of the two workloads in page-level. We assume that the page size of NAND flash is 2KB. Figure 3(a) and Figure 3(b) show page-wide view of data pattern of each workload for SLC. As previously described, the probability for each bit to be programmed is remarkably different and it is neccessary to be wear-leveled. The reason why the counts of lower indexed bytes in a page are slightly higher than the others in case of Linux source codes, is beacause the size of many files in Linux source tree is lower than the page size and we obtained the data by only accumulating the count of bit value having logic 0. In case of smaller sized file than the size of one page, the dummy data contents in higher offset than the size of files, had better being filled with logic '1' for NAND flash's endurance.

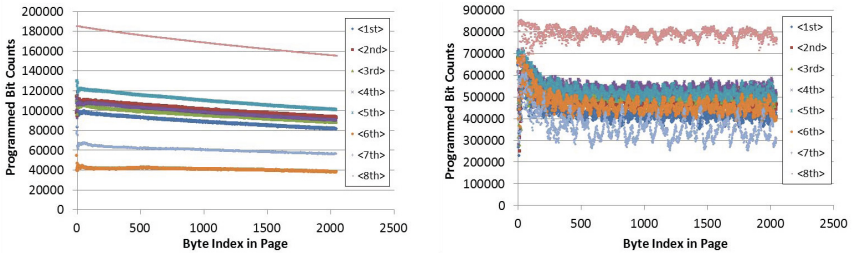


Fig. 3. (a) Page-wide view of wear-out for Linux source code, (b) Page-wide view of wear-out for DB workload

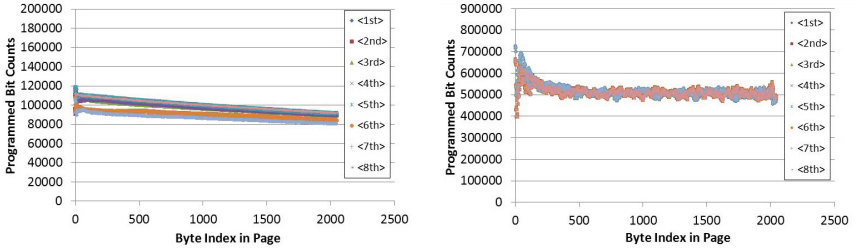


Fig. 4. (a) Results of our wear-leveling scheme for Linux source code, (b) Results of our wear-leveling scheme for DB workload

The results in Figure 4(a) and Figure 4(b) are obtained by applying our wear-leveling scheme, bit-level shifting, on the workloads in case of SLC. By applying the proposed scheme, the estimiated wear-out counts of each cell become nearly uniform. With comparing the maximum count, in case Linux source code, the degree of wear-out becomes 36% lower and 15% lower in case of DB workload. Additionally, in case of DB workloads, lower offset of bytes in a page tends to be worn out more than the others in spite of bit-level wear leveling.

5 Conclusion

We introduce our observation that bit-level data patterns of widely-used file data are not uniform and, based on the breakdown mechanism of oxide, we also propose simple wear-out models for SLC NAND flash. To expand endurance of NAND flash, we emphasize the consideration of bit-level wear-leveling scheme and proposed simple approaches. We expect that our solutions can be adopted in the controller-level togther with other wear-leveling scheme or other encoding schemes.

References

1. DBT2 benchmark, http://sourceforge.net/apps/mediawiki/osdldbt/index.php
2. Bin, Z.W., Ming, X., Zheng, X., Kang, Z.A.: Charge-to-breakdown (Q_{BD}): a method to monitor the ultrathin tunnel oxide in E^2PROM. In: Proceedings of the 1998 5th International Conference on Solid-State and Integrated Circuit Technology, pp. 287–290 (1998)
3. Chang, L.-P., Huang, L.-C.: A low-cost wear-leveling algorithm for block-mapping solid-state disks. SIGPLAN Not. 46(5), 31–40 (2011)
4. Chang, L.-P., Kuo, T.-W.: Efficient management for large-scale flash-memory storage systems with resource conservation. TOS 1(4), 381–418 (2005)
5. Chang, Y.-H., Hsieh, J.-W., Kuo, T.-W.: Improving flash wear-leveling by proactively moving static data. IEEE Trans. Computers 59(1), 53–65 (2010)
6. Hussain, S.A., Mansoor, A.: Flash modelling for wearleveling algorithms. In: High Capacity Optical Networks and Enabling Technologies (HONET), pp. 267–272 (December 2011)
7. Murugan, M., Du, D.H.-C.: Rejuvenator: A static wear leveling algorithm for NAND flash memory with minimized overhead. In: Brinkmann, A., Pease, D. (eds.) MSST, pp. 1–12. IEEE Computer Society (2011)
8. SQLite. Sqlite official home page, http://sqlite.org

HASV: Hadoop-Based NGS Analyzer for Predicting Genomic Structure Variations

Gunhwan Ko[1], Jongcheol Yoon[1], and Kyongseok Park[2,*]

[1] Korean Bioinformatoin Center (KOBIC), Daejeon, South Korea
{kogun82,yescheol}@kribb.re.kr
[2] Korea Institute of Science and Technology Information (KISTI), Daejeon, South Korea
gspark@kisti.re.kr

Abstract. The NGS technology produces large scale biologic data sets much cheaper and faster than the previous methods. As it is almost impossible to store or analyze such large scale NGS data with a traditional method on a commodity server, many problems arise. Hadoop is an alternative to this requirement. We aim to address the issues involved in the large scale data analysis on the cloud in bioinformatics. Accordingly, we propose analysis service for predicting genome structural variations associated with diseases by using Hadoop. The result of this study reveals that the system proposed in this study efficiently predicts genomic variations from large scale data sets.

1 Introduction

With the current NGS (next generation sequencing) [1] technology, it is possible to analyze one person's genome within about a week. The data volume produced from NGS technology currently ranges from a few TB to PB scale. It is almost impossible to store or analyze this voluminous NGS data in a single commodity machine: expensive hardware is required for data analysis; existing frames are not enough to fully manage resources; managing such data requires huge storages; and lots of computing resources are required to analyze voluminous NGS data. In the field of bioinformatics, many analysis services are provided by using Hadoop to address the issues involved in biologic data analysis. We propose a Hadoop-based algorithm HAVS for predicting genomic variations and related to diseases. The algorithm covers the structural variation types such as translocation, inversion, insertion, and deletion. We also prove that our algorithm shows better performance and accuracy than BreakDancer [2]. The experiments are performed on Hadoop clusters with the proposed algorithm and the YRI sequencing data provided by the 1000 Genome Project [3].

2 Detecting Genome Structural Variation of Cloud Scale

2.1 PEM Based Approach and Genome Structural Variation Region

Typical PEM-based structural variation detection methods use paired-end reads. Two paired reads created in a genome (case) to be detected have distance information each

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_49,

other. When two reads are mapped with a reference genome of which the sequence is already known, structural variation is detected by computing the difference between the distance mapped with the reference genome and the distance in the case. In this case, because the read is mapped with the reference genome in consideration of both forward direction and backward direction, it is possible to detect inversion. The PEM-based methods of finding and analyzing paired reads support resolution even higher than the array-based methods. The PEM-based structural variation detection methods analyze and characterize the mapped type of two reads. Here, the characteristics are referred to as an event or signature. The signatures are categorized into insertion, deletion, inversion, linking, and duplication depending on their nature and mapping types. After reads are mapped and signatures are found, clustering is then carried out to find a region where signatures are crowded. Clustering is used to effectively filter candidate regions of actual structural variation, rather than computing positions where structural variation occurred by using one signature [4]. That is, clustering contributes to improving prediction reliability by removing accidental matching parts, and to accurate prediction of the positions of structural variation. After clustering, signatures mapped with one cluster region are used to compute a structural variation region to represent the relevant cluster. In this case, both ends where variation occurred are called a breakpoint, to compute the method of determining signatures which consist of a cluster, and an actual breakpoint.

2.2 Extracting Genome Structural Variation Based on Statistics

Extraction of genome structural variation based on Hadoop is carried out in a 3-step MapReduce workflow. The first step is to classify the sequences as normal sequences (concordant), of which the direction of paired-end reads is correct and in which the mapping distance belongs to the range of ±2.7SD standard deviation of the mean fragment size of the reads, and the sequences that do not belong to the range as variation (discordant). The insert size, the mean fragment size, and standard deviation are computed, to be used as a reference for normal sequences among the classified sequences. And the sequences predicted to be structural variation are classified for each type. The second analysis process is to compute the mean insert size of sequences by structural variation types for only the sequences regarded as structural variation. The third analysis step is to cluster the structural variation sequences on the basis of the information computed in the above analysis process, and to predict a breakpoint in which structural variation occurred.

2.2.1 Filtering Paired-End Reads and Computing Standard Deviation

On the basis of the mapped paired-end reads, normal sequences and the sequences predicted to be structural variation are classified. In this case, the reference for classifying structural variation is defined in HASV as described below. For the inferred insert size, only positive numbers can be used in order to avoid overlapping sequences, and the sequences are regarded as a structural variation sequence if it is a negative number. If the mapping quality is not greater than the minimum quality, it is regarded as a structural variation sequence. If the maximum fragment size is above the allowable range, it is decided as a structural variation sequence. If the paired-end reads do not exist in the same chromosome, it is specified as a structural variation sequence.

If a sequence is not mapped with the reference sequence, it is specified as a structural variation sequence.

After classification on the basis of the classification conditions for each structural variation type as described above, the structural variation types are given a relevant id as a key. More efficient and faster analysis can be implemented by gathering the same structural variation types to form clusters on the basis of the key in the step of detecting and analyzing genome structural variation. The classified normal sequences are used for detecting the genome structural variation by finding the mean read length and the mean fragment size of normal sequences. The following equation 1 is used for the computation.

$$\bar{m}(mean\ fragment\ size) = \frac{\sum_{i=1}^{N} x_i(fragment\ size)}{N(concordant\ rad\ count)}$$
$$\bar{r}(mean\ read\ length) = \frac{\sum_{j=1}^{J} y_j(read\ length)}{J(read\ count)} \quad (1)$$

For the fragment size of sequences predicted to be structural variation, if the mean fragment size computed by means of the above equation 1 is $\bar{m}$, the maximum threshold is computed by using the fragment size of the corresponding sequences for $\bar{m} < \text{fragment size}$. For $\bar{m} > \text{fragment size}$, the minimum threshold is computed. The following equation 2 is used to compute each threshold.

$$u_k > \bar{m}\ then\ MAX : U = \bar{m} + \left(\sqrt{\sum_{k=1}^{K} \frac{(u_k(long\ fragment\ size) - \bar{m})^2}{K(long\ fragment\ count) - 1}} \times C \right)$$
$$l_p > \bar{m}\ then\ MIN : L = \bar{m} - \left(\sqrt{\sum_{p=1}^{P} \frac{\left(l_p(short\ fragment\ size) - \bar{m}\right)^2}{P(short\ fragment\ count) - 1}} \times C \right) \quad (2)$$

The parameter C used in the above equation 2 is a cutoff value. Because HASV uses all paired-end sequences predicted to be structural variation for analysis, the number of occurring errors also increases with the increasing number of sequences to be detected. As a result, the cutoff value is used to select significant structural variation sequences in terms of probability to decrease false positive. Unlike typical structural variation detection tools which carry out analysis on the basis of statistics derived through random sampling, HASV uses entire data and can thus derive more accurate statistical information. Furthermore, accurate statistical information contributes to improving specificity and accuracy in detecting structural variation.

2.2.2 Computing Insert Size for Each Structural Variation Type

Because the insert size between the reads in the paired-end reads of the predicted structural variation is different by types, the mean insert size for each structural

variation is computed in addition to the second filtering process by using the statistical information derived from normal paired-end reads. If the fragment size(x_i) is $x_i < (\bar{m} - \sigma \times C)$ or $(\bar{m} + \sigma \times C) < x_i$ while the number (n) for each type is $n > 0$ for the paired-end reads predicted to be structural variation, the mean insert size $\bar{d}$ of the paired-end reads for each structural variation type is defined as described below.

$$\bar{d} = \frac{\sum_{i=1}^{n}(fragment\ size_i - read\ length \times 2)}{n(structural\ variation\ count)}$$

If the insert size $\bar{d}$ for each computed structural variation type is $\bar{d} \leq 0$, $\bar{d} = \bar{m} - \bar{r} \times 2$ is substituted for the value of $\bar{d}$ to be used for analysis.

2.2.3 Algorithm for Detecting Genome Structural Variation Sequence

The Hadoop-based algorithm for detecting genome structural variation is composed of the following two stages. The first stage uses the statistical information computed through the previous analysis step to cluster sequences in the location similar to the reference genome. The second stage computes the start position and the end position where structural variation occurred from a representative sequence predicted by clustering in the first stage to predict the breakpoint which is a section with variation in the entire sequences. The genome structural variation region is then detected to carry out the merging process. The process is carried out for the sequences classified as structural variation through the previous validation analysis process in order to cluster structural variation sequences mapped in similar locations in the reference genome. It is better than using only signatures found through one read mapping to compute the location with structural variation.

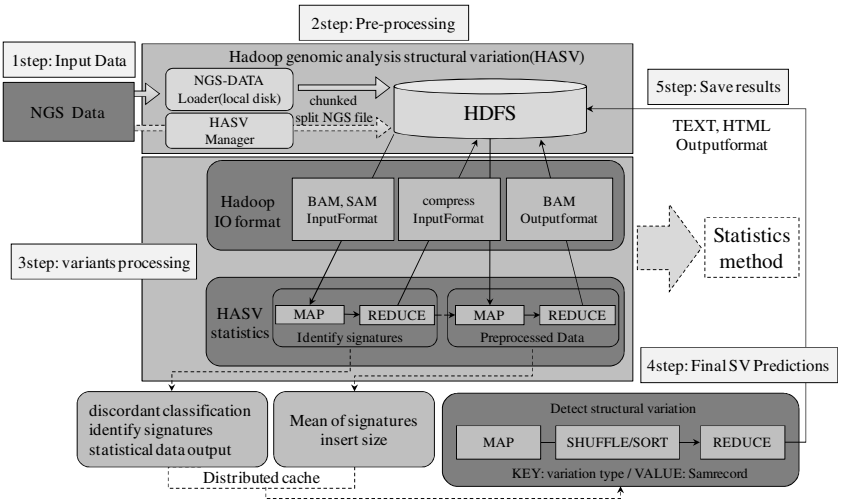


Fig. 1. Data flow of HASV

The MAP step of the structural variation detection algorithm receives the sequence information predicted to be a structural variation paired-end read. The KEY for combining a signature with the read ID becomes K1, and SAMRecord with the paired-end sequence information becomes VALUE V1. The paired-end read predicted to be structural variation and the statistics are received by means of the config object to use them for computing the breakpoint of the paired-end reads predicted to be structural variation. If the paired-end read predicted to be structural variation is x_i, x_{i-1} and x_i are positioned in the same chromosome.

If the sorting start position of x_{i-1} is p1, and the sorting start position of x_i is p2, p2- p1<minimum region length. If the number of sequences which form one cluster is smaller than 100, they are regarded as paired-end reads in the same cluster. The minimum length region is a value used in filtering paired-end reads mapped abnormally in the clustering step, and contributes to improving the specificity of detected structural variation sequences. The paired-end read sequences in a cluster are used to select a representative sequence on the basis of the sorting start position and the sorting end position, in order to predict the breakpoint of variation.

3 Experiment Result and Analysis

3.1 Experiment Environment and Data

In this experiment, the accuracy of detecting structural variation of the HASV algorithm was verified and the analysis speed improvement index was then evaluated with respect to the increasing number of nodes in the Hadoop-based cluster. The experiment data was the UCSC hg18 human genome reference sequence and the sequence data of YRI provided in the 1000 genome project. The sequence sorting program was BWA 0.5.9 [5]. The error range from the actual region of structural variation was measured by comparing the structural variation region detected by the HASV algorithm suggested to verify the structural variation detection accuracy of the algorithm with the structural variation region reported to the DGV [6]. The structural variation database of DGV has detailed records of human structural variation found through all sorts of wet/dry experiments. It is possible to search regional or personal data.

3.2 Experiment Result and Analysis

3.2.1 Measuring Accuracy of HASV Algorithm

While using the same analysis method as HASV, a comparison and an analysis was made with/of the result of BreakDancer which is widely used for structural variation, in order to verify the accuracy of the HASV algorithm. The accuracy of structural variation detection was verified in the YRI chromosome 10 by using HSAV and BreakDancer. The actual deletion structural variation discovered through the DGV experiment occurred in the region 490486 ~ 490606 of the chromosome 10 in the YRI sample data. While BreakDancer predicts the deletion structural variation in the region 489751 ~ 490733, HASV predicts deletion structural variation in the region 490491 ~ 490625. The experiment result reveals that HASV predicts structural variation more accurately than BreakDancer. And it is reported that deletion structural variation exists

through 2080877 ~ 2081202 in DGV. It is shown that BreakDancer does not predict deletion structural variation. However, although HASV has an error, it predicts deletion structural variation in the region 2081058 ~ 2081214.

3.2.2 Evaluating HASV Analysis Speed in Hadoop-Based Cluster

In this experiment, the effect of performance improvement was evaluated with respect to the increasing number of nodes in the Hadoop-based. The experiment result reveals linear improvement with respect to the increasing number of nodes in the overall analysis performance. This is a result different from the general theory that performance improvement is not doubled due to data exchange between input/output to/from nodes and the nodes although the number of nodes in a distributed system is doubled.

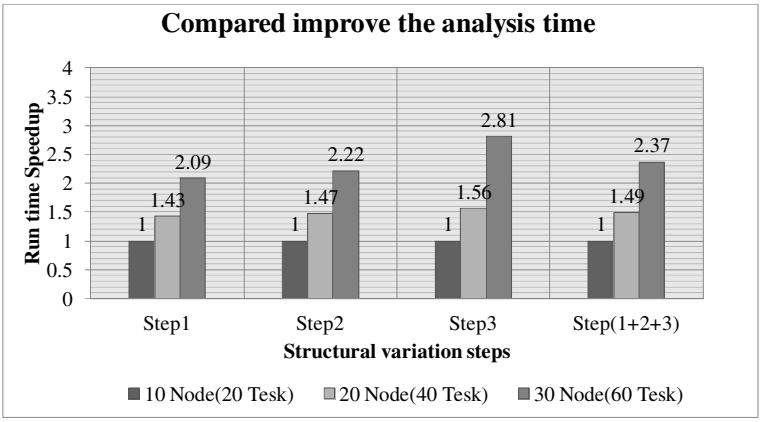


Fig. 2. Analysis performance improvement ratio with respect to the increasing number of nodes

From this experiment, the measurement is that performance improvement effect with respect to the increasing number of nodes was maximum 70 ~ 75% although there might be some difference depending on executed applications and the relation with MapReduce. For example, this means that, if 10 additional nodes are added, the performance is not 10 times faster, but maximum 7 ~7.5 times faster, in comparison with execution in one node. Because the aforementioned performance improvement value linearly increases, it is possible to improve analysis time by adding nodes without changing programs. Therefore, Hadoop is useful in bioinformatics which has a lot of data to be processed and tasks to be divided into many steps.

4 Conclusion and Future Studies

This study proves that it is possible to improve performance in a cloud computing environment when analyzing big NGS data, and to detect and analyze structural variation by using the big NGS data based on the MapReduce programming model. That is, it is possible to address limited performance due to insufficient computing resources which is an issue in analyzing big data in bioinformatics, in the scale-out

method which is an advantage of Hadoop. However, this structural variation detection analysis is just for chr10 of the YRI sample data, and the result is not for full applications. For future studies, it is necessary to detect structural variation for whole genomes by using HASV, to compare the variation result of different structural variation detection programs and to verify the accuracy. It is also necessary to design a system useful for big data analysis in bioinformatics by increasing performance which is a current issue in Hadoop.

References

1. Xia, J., Wang, Q., Jia, P., Wang, B., Pao, W., Zhao, Z.: NGS catalog: A database of next generation sequencing studies in humans. Hum. Mutat. 33, E2341–E2355 (2012)
2. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K., Ding, L., Mardis, E.R.: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6, 677–681 (2009)
3. Kuehn, B.M.: 1000 Genomes Project promises closer look at variation in human genome. JAMA 300, 2715 (2008)
4. Medvedev, P., Stanciu, M., Brudno, M.: Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods 6, S13–S20 (2009)
5. Li, H., Durbin, R.: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595 (2010)
6. Duclos, A., Charbonnier, F., Chambon, P., Latouche, J.B., Blavier, A., Redon, R., Frebourg, T., Flaman, J.M.: Pitfalls in the use of DGV for CNV interpretation. Am. J. Med. Genet. A 155A, 2593–2596 (2011)

Provisioning On-Demand HLA/RTI Simulation Environment on Cloud for Distributed-Parallel Computer Simulations

In-Yong Jung, Byong-John Han, and Chang-Sung Jeong*

Department of Electrical Engineering, Korea University, Seoul, Korea
{dekarno,guru1013,csjeong}@korea.ac.kr

Abstract. In distributed parallel computer simulation, there are various simulation platforms supporting HLA on the Grid or Cloud, but almost of them shows lack of various aspects because of limitation of HLA. In this paper, we present an architecture of Cloud Distributed-Parallel Simulation Platform for HLA (CDSPH) supporting distributed parallel computer simulation on the cloud. It offers on-demand resource provisioning for scalable simulation, self-organization of adaptive simulation environment, and enhanced security by isolation between federation executions. Besides, it support web-based user interface to support easy access and simulation management.

Keywords: Cloud Computing, PaaS, Distributed-parallel simulation, HLA, RTI.

1 Introduction

The HLA (High Level Architecture)[1] is an architecture which purpose interoperability among entities and reusability of codes for distributed-parallel computer simulations. Simulation entity is defined as federate for each, and RTI (Runtime Infrastructure) coordinates operations and data exchange between federates during a runtime execution. HLA is defined under IEEE Standard 1516 and used to support data analysis, engineering and training in a number of different domains such as aerospace industries, military, traffic controls and manufacturing. However, it shows lack of efficient allocation of simulation resources, automatic self-organization and security.[2] There are various approaches to overcome its weak points by integrating and launching HLA simulations on the Grid or Cloud environments.[2-3] However, Grid simulation frameworks should be strengthen a number of aspects such as heterogeneity of its resource pool, multi-tenancy and fault tolerance.[3] Besides, a number of researches presented cloud simulation middleware architecture adapting virtualization technology, but their embed cloud RTI service is centralized which can be a bottleneck of the interaction between hosted simulation entities. It is necessary to design a platform architecture based on HLA for scalable distributed-parallel

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_50,

simulation with automated resource provisioning and federation execution with dedicated RTI service.

Cloud computing is getting spotlight as the main key words for IT industry today. It can integrate various IT technologies that supply on-demand resources, useful platforms and applications, so it is becoming useful solution for cost reduction, self-organizing and flexible scale extension.[4] Therefore, cloud technologies can suggest an easier way for launching user's simulation entities through integrating simulation infrastructure with scalable resource provisioning, automated software allocation and execution. In this paper, we present an architecture of Cloud Distributed-Parallel Simulation Platform for HLA (CDSPH) supporting distributed-parallel computer simulation. It offers a virtual environment which provides automated resource provisioning, allocating and execution of HLA federations on the cloud. It makes up for the weak points of HLA and offers easy way to operate and manage simulations to the user.

2 Cloud Distributed-Parallel Simulation Platform for HLA

In this section, we describe several features of CDSPH, and then its architecture in detail. It builds virtual simulation environment on the cloud VMs automatically, and supports user to launch their HLA federation easily. We describe key features of CDSPH as follows.

1) On-demand resource provisioning for scalable distributed-parallel simulation: CDSPH has IaaS management layer to access and control cloud infrastructures for supporting on-demand resource provisioning. When user wants to launch his federation on the cloud, CDSPH provides adequate VMs on its managed cloud infrastructure dynamically. Lifecycle of VM is managed by cloud infrastructure, user needs not to manage about each VMs
2) Self-organized simulation environment via adaptive platform provisioning: CDSPH configures appropriate simulation environment on the VMs for federation execution. Each federates of the federation and dedicated RTI instance are deployed on each VMs. Deployed federation is launched and monitored by Resource agent on each VM automatically.
3) Enhanced security through dedicated resource and RTI instance: Via CDSPH, execution of a federation gets dedicated VM resources and RTI instance which are not shared with the other federations. This isolation gives enhanced security level of launching users HLA simulations.
4) Easy access and interaction via web-based user interface: Through web-based Cloud Simulation Portal (CSP), user can manage and interact with his federation execution launched on the clouds.

CDSPH consists of Cloud Simulation Portal (CSP) and Cloud Distributed-Parallel Simulation Framework (CDSF) as shown in Fig. 1.

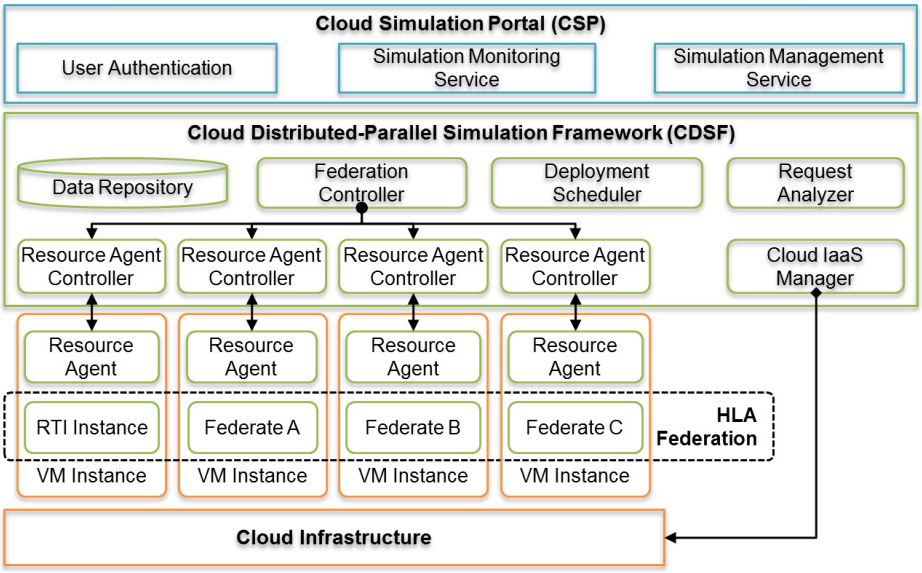


Fig. 1. The architecture of CDSPH with cloud infrastructure

Cloud Simulation Portal (CSP) offers a web-based graphical user interface to support request of launching federation, simulation monitoring and user authentication. User chooses appropriate simulation environment and put simulation configurations such as federation object model (FOM) described in FDD written in a Lisp-like syntax for HLA v1.3, or XML for IEEE Standard 1516.[1] RTI-specific configurations which user wants are customized via CSP, too. After collecting user's requirements, CSP submits users request named Simulation Description (SD) to the CDSF to launch user's federation.

Cloud Distributed-Parallel Simulation Framework (CDSF) composes a software environment for launching HLA federates. It adopts agent-based resource management scheme[4], to allocate, launch, and monitor various federates on the VMs. When user request is arrived, Request Analyzer reads SD submitted by CSP, and interprets user's requirement. Then CDSF creates new Federation Controller and requests VMs to the cloud infrastructure via Cloud IaaS Manager, the abstracted IaaS API component. After new VMs are created, pre-installed program named Resource Agent on each VM starts automatically, and tries to connect to the CDSF. CDSF accepts each access by creating new Resource Agent Controller one by one. If all VMs are created and successfully connected, Federation Controller determines roles of each VMs via Deployment Scheduler. Therefore, each federate and RTI instance are allocated to the suitable VM. Because HLA is architecture, not on-the-wire protocol, participants in a federation can be dependent to a specific RTI by using its libraries. Therefore, well-suited RTI must be selected for federation execution. Federation Controller sends software such as federates, RTI, user configuration data and commands. After preparation of launching federation and RTI startup, each Resource Agents executes assigned federates on their VM. All simulation process and

logs are monitored by Resource Agents and integrated by Federation Controller, and this information is offered to the CSP. Therefore, users can check status of simulation process via CSP.

3 Experiments

We implemented prototype of CDSPH in java 6, and performed two kinds of experiments for evaluation. CDSPH is connected to a Eucalyptus cloud infrastructure based on Ubuntu Enterprise Cloud.[5] Our cloud infrastructure consists of 1 Cloud/Cluster controller node and 4 worker nodes with KVM hypervisor. It was deployed on servers with 2 Xeon E5606 2.13GHz processors and 24GB memory. Tested VM image is based on Ubuntu 10.04 LTS 64bit server, and includes pre-installed Resource Agent, java 6 and cloud-init.

The first experiment is on measuring provisioning time of VMs for HLA simulation. Fig 2 shows the cumulative average provisioning time of each stage: (a) VM Creation, (b) Data Distribution for transfering user federate program, and (c) Federation Scheduling for deploying RTI and user federates. Instance type of created VMs has dual core processors and 4GB memory. According to Fig. 2, VM creation time is increased with a number of VMs. Time for VM creation is the greatest portion, so total provisioning time is increased, too. Because our cloud infrastructure has 4 worker nodes and its scheduling policy is round-robin, time is significantly increased when the number of VMs is over each multiples of four.

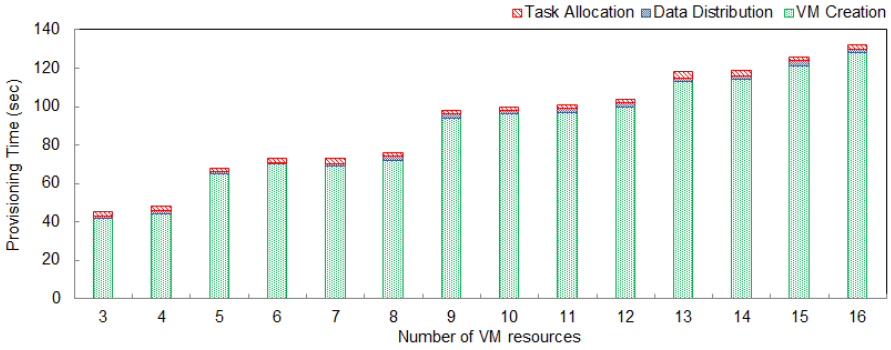


Fig. 2. The cumulative provisioning time of VM resource to allocate federation

The second experiment is on measuring performance of federation on two kinds of simulation environments. We executed portable RTI and a sample federation [6] on the cloud environment provided by CDSPH and native environment. Federation consists of 5 federates; Manager, Production, Transport, Consumption, and Viewer. This time-constraint federation simulates a simple restaurant, and includes various HLA behaviors such as publishing and subscribing their attributes. Its time interval is 1 second, and we executed the federation during 1000 time advances. In each

advances, federates finishes their task and wait until the next advance arrives. Therefore, we can measure the performance of simulation by measuring the length of idle time (sleeping time). We deployed RTI and each federates on six VMs on cloud, and six native servers which have Core 2 Duo E6550 2.33GHz processor and 4GB memory for each. Fig. 3 shows the idle time of federation in each intervals. According Fig. 3, we can compare the performance of federation execution launched on two simulation environments. The idle time of federation execution on cloud is about 30ms less than that on native cluster environment. Table 1 shows statistics of experiment. Variance of federation execution on cloud is about 5 times higher than that of native environment, and its fluctuation is 60% higher, too. It means that performance on cloud needs to be improved and stabilized in further research. However, these results show that federation execution on cloud achieves about 97% of performance of native environment.

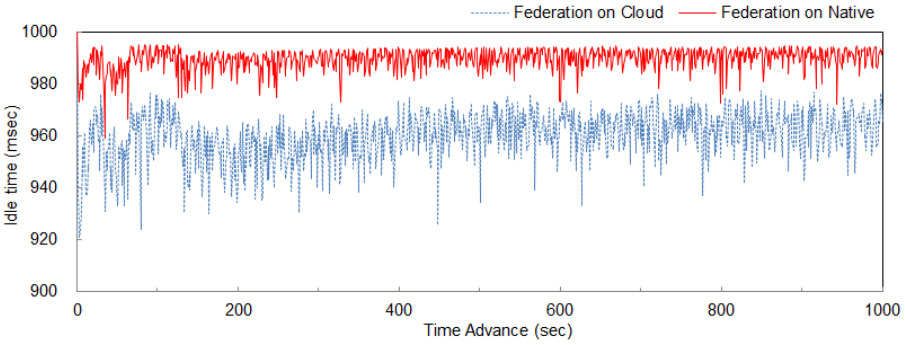


Fig. 3. The comparison of idle time of the federation execution on simulation environments

Table 1. The comparison of experiment statistics

	Max. (ms)	Min. (ms)	Average (ms)	Variance
Federation on Native	995	959	990.1	18.0
Federation on Cloud	979	921	960.8	85.8

4 Conclusion and Future Works

This paper has presented Cloud Distributed-Parallel Simulation Platform for HLA (CDSPH) for distributed parallel simulations on the cloud environments. It provides adequate on-demand VM resources for scalable simulation, and adaptive simulation environment configuration for self-organization of federation execution. Isolation of federation executions gives enhanced security and resolves bottleneck of centralized RTI services. Besides, it support web-based user interface to support easy access, monitoring and simulation management. We are planning design and development of

enhanced user interface, enhanced federate scheduling algorithm, auto scaling and fault-tolerant architecture as a future works.

Acknowledgments. This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA (National IT Industry Promotion Agency), Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0006425), and Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program (R2012030096).

References

1. High Level Architecture, http://en.wikipedia.org/wiki/High-Level_Architecture
2. He, H., Li, R., Dong, X., Zhang, Z., Han, H.: An Efficient and Secure Cloud-Based Distributed Simulation System. Appl. Math. Inf. Sci. 6(3), 729–736 (2012)
3. Li, B., Chai, X., Hou, B., Li, T., Zhang, Y., Yu, H., Han, J., Di, Y., Huang, J., Song, C., Tang, Z., Wang, P., Shi, G., Wang, X.: Networked modeling & simulation platform based on concept of cloud computing-cloud simulation platform. J. Syst. Simulat. 21, 5292–5299 (2009)
4. Jung, I., Lee, D., Han, B., Kim, K., Jeong, C.: Social Media Cloud Platform for Distributed-Parallel Data Processing and Infrastructure Management. In: Proceedings of the 3rd International Conference on Internet (ICONI 2011), pp. 607–611 (2011)
5. Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., Zagorodnov, D.: The Eucalyptus Open-Source Cloud-Computing System. In: Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 124–131. IEEE Computer Society, Los Alamitos (2009)
6. Kuhl, F., Weatherly, R., Dahmann, J.: Creating computer simulation systems: An introduction to High Level Architecture. Prentice Hall PTR, Upper Saddle River (2000)

Fast Shear Skew Warp Volume Rendering Using GPGPU for Cloud 3D Visualization

Chang-Woo Cho[1], Ki-Hyun Kim[2], Ki-Young Choi[2], and Chang-Sung Jeong[2,*]

[1] Division of Internet & Multimedia Engineering, Korea University, Seoul, Korea
chocwjc@korea.ac.kr
[2] School of Electrical Engineering, Korea University, Seoul, Korea
{cocoball,2xx195,csjeong}@korea.ac.kr

Abstract. We present a method for fast volume rendering using GPGPU for Cloud 3D Visualization. Each of the threads is processed in parallel on the GPU. Our algorithm is to use the computational power of graphics processors to speed up the rendering process of the Shear-Warp algorithm. Our GPU-based method provides real-time frame rates and outperforms the GPU-based implementation. Our experimental results show that our algorithm is much faster than the CPU-based Shear-Warp volume visualization in terms of rendering speed and image quality. It can be launched on GPU computing clusters provided by cloud infrastructures such as Amazon EC2.

Keywords: Shear warp volume rendering, Cloud 3D Visualization, Parallel Processing.

1 Introduction

Parallel computing is required expensive multi-core CPU or Multiple computers, even though only a few years ago. However, parallel computing can be realized easily by GPGPU technology. Also, volume rendering [1, 2] is a technique for visualizing 3D volume data from CT or MRI on the screen. The shear warp algorithm [3] is one of the fastest algorithms among direct volume rendering. However, the existing shear warp algorithm has some defects such as increases in memory consumption and preprocessing time as well as deterioration in image quality. Especially, for collaborative visualization for medical application, the importance of real time rendering of MRI or CT data becomes increasing for the fast diagnosis of the patient status in remote sites.

In this paper, we propose a GPU-based implementation of the shear warp algorithm for cloud 3D volume rendering applications. To our knowledge, it is the first implementation on the GPU. And then, we compare it to a CPU-based shear warp algorithm and to a GPU-based implementation in terms of flexibility and frame rate. It can be launched on GPU computing clusters provided by cloud infrastructures such as Amazon EC2.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_51,

2 Shear Skew Warp Volume Sketch Algorithm on GPU

In this section, we first describe the algorithm with its acceleration schemes, and then present the details of our algorithm using GPGPU.

2.1 Sampling of Volume Data by GPU

As in Fig.1, an intermediate image is obtained from the sampled volume data by shear skew transformation. Before the original volume data is allocated to the fixed memory in host memory, the sampling of volume data is copied onto the texture memory of GPU device. Our algorithm has improved the repetitive shear-skew and project into an intermediate image by the texture memory. It has speed up about 20% on average. The values of x, y, z axis to move the units are copied onto the constant memory of the GPU, so bank conflict decreased when the constant memory is simultaneously accessed. And then, the final image is converted from the intermediate image by the warping and up-scaling process using GPGPU, which will be described in details in the next subsection.

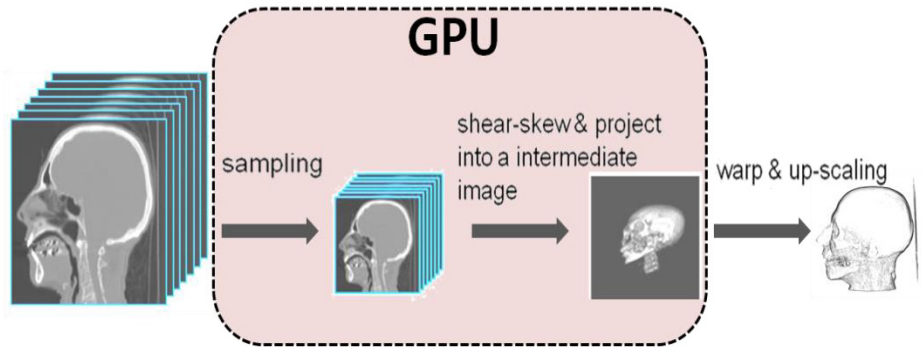


Fig. 1. Overview of shear skew warp volume sketch algorithm

2.2 Implementation of Shear Warp Algorithm on GPU

Our aim is to use the computational power of grapics processors to speed up the rendering process of the Shear-Warp algorithm The power of the graphics processors comes from a highly parallel structure using built-in mathematical operations such as matrix product and texture filtering. So our approach uses the process of resamping and compositing on GPU kernel fuction using CUDA. As shown in Fig.2, we determine the number of threads and blocks that one thread can be responsible for one pixel of a volume slice. And each of the threads is processed in parallel on the GPU. For example, 32 x 32 blocks (256 thread per a block) are needed to process a 512 x 512 size of volume slice at a time. Although three loop computations are needed to resample and composit values of all voxels in x, y and z directions on CPU, one loop computation is needed with the help of GPU.

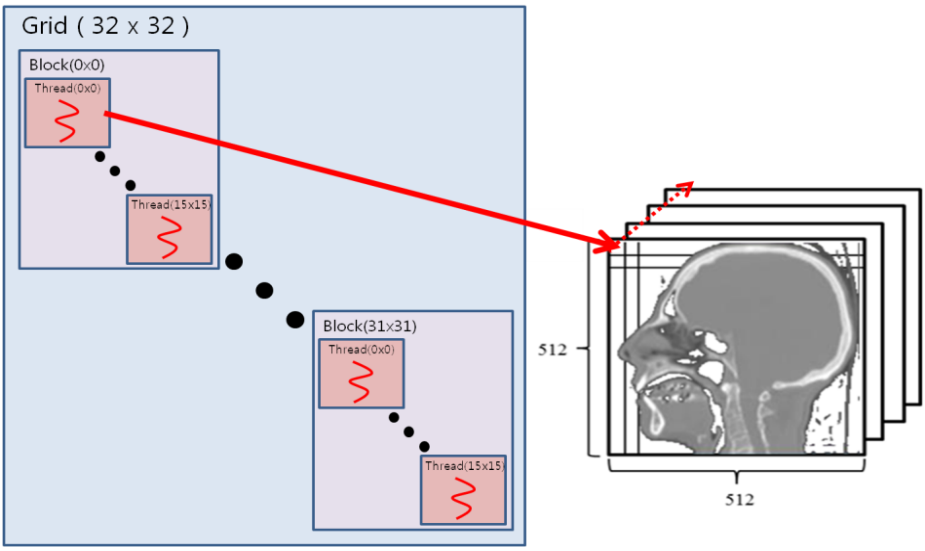


Fig. 2. One thread is responsible for one pixel of a volume slice for resampling and compositing

3 Performance Evaluation

In this section, we describe the performance evaluation of our algorithm. It is implemented and tested for eight kinds of volume data in PC with Intel Core2 Duo 2.66GHz CPU, GeForce GTX560, 2G RAM and Windows XP OS. We compare the performance for the previous shear skew warp[4] and our shear skew warp sketch algorithms in terms of rendering speed for six different size of volume data as shown in Table.1.

Table 1. Eight different sizes of volume data

Data	Dimension	bits/voxel	Size(Mbytes)
engine	256x256x110	8	7.208960
CT head	256x256x225	8	14.745600
foot	256x256x256	8	16.777216
stag beetle	416x416x247	8	42.744832
CTA Brain	512x512x279	8	73.138176
bunny	512x512x361	8	94.633984

In Fig. 3, (a), (b), (c), (d), (e), (f) is the resulting images rendered by shear skew warp volume sketch algorithm using GPGPU.

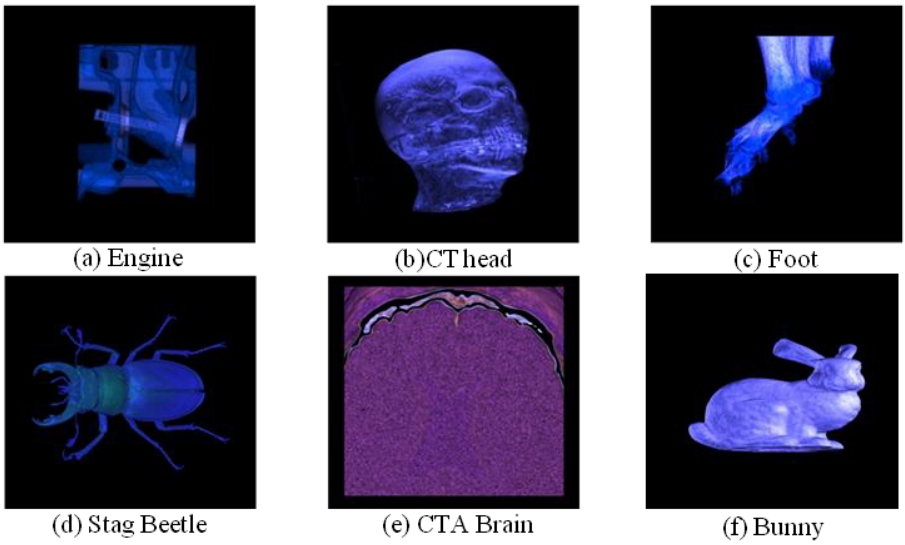


Fig. 3. Resulting images by shear skew warp volume sketch algorithm using GPGPU

3.1 Rendering Time Using GPGPU

Fig. 4 shows the comparison of rendering speed among two algorithms and using GPGPU. Our algorithm has speed up about 50% on the average with minimum 1.94 and maximum 3.01 times compared to the previous one due to the reduction of samples through Poisson disk sampling. Also, Based GPGPU has speed up about 15% on the average. It shows that the preprocessing time and speed up ratio is almost proportional to the size of volume data. Using GPGPU, first of all necessary data transfer to GPU device memory, then the calculated results are transmitted to the CPU host memory again. So, it shows that the result of lower than the ideal value.

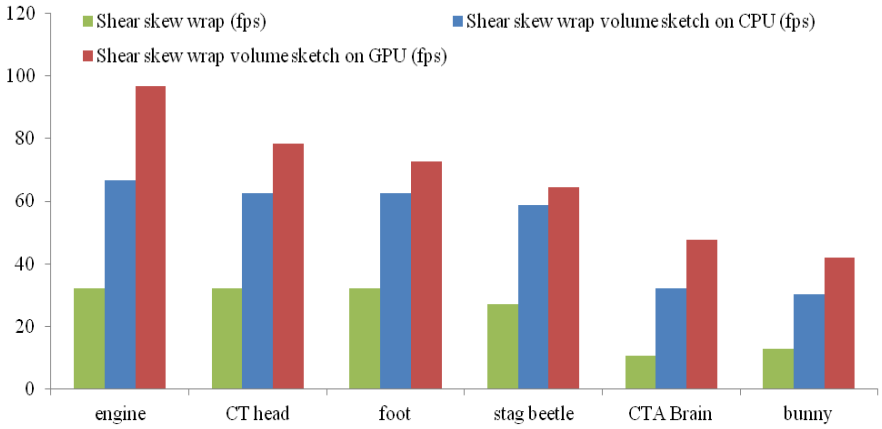


Fig. 4. Comparison of rendering speed

4 Conclusions and Future Works

In this paper, we have presented a method for fast volume rendering on the GPU using CUDA. The GPU-based shear skew warp volume visualization is processed in parallel. And our algorithm is to use the computational power of graphics processors to speed up the rendering process of the shear warp algorithm. The method provides real-time frame rates and outperforms the GPU-based implementation. Our experimental results show that our algorithm is much faster than the CPU-based Shear-Warp volume visualization in terms of rendering speed and image quality. It is launched on GPU computing clusters provided by cloud infrastructures such as Amazon EC2. We are planning optimization and development of the algorithm as a future works.

Acknowledgments. This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA (National IT Industry Promotion Agency), Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0006425), and Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program (R2012030096).

References

1. Drebin, R., Carpenter, L., Hanrahan, P.: Volume Rendering. J. Computer Graphics 22, 65–74 (1988)
2. Levoy, M.: Display of surfaces from volume data. IEEE Computer Graphics & Applications 8, 29–37 (1988)
3. Lacroute, P., Levoy, M.: Fast volume rendering using a shear-warp factorization of the viewing transformation. In: Proceedings of SIGGRAPH, Orlando, Florida, pp. 451–458 (1994)
4. Kiyoung, C., Sungup, J., Changsung, J.: CPU-Based Speed Acceleration Techniques for Shear Warp Volume Rendering. Multimedia Tools and Applications (2012) (published online)

A Vision-Based Robust Hovering Control System for UAV

Tyan Vladimir, Dongwoon Jeon, and Doo-Hyun Kim*

Konkuk University, Department of Internet & Engineering,
Seoul, Republic of Korea
{auron,dongwoon,doohyun}@konkuk.ac.kr

Abstract. This paper introduces an algorithm for real-time line detection and tracking utilizing the Graphic Processing Units (GPUs) for UAV's vision-based hovering control system. We concentrate that there are many of lines where UAV can fly, and extract meaningful line to grasp of vehicle's attitude. We implemented image processing techniques on GPUs for real-time performance because detection and tracking of lines need huge computational resources. Experiments show affordable frame throughput that our approach is feasible in real flight.

Keywords: CUDA, Line Detection, Line Tracking, Hough Transform, Hovering System.

1 Introduction

Recently, there are a lot of software programs and hardware dealing with image or video processing in real-time. These technologies find use in a wide range of applications: medical computer vision, industry automation, military applications, robotics and etc. One of the newest application areas is autonomous vehicles, which include submersibles, ground vehicles, aerial vehicles, and Unmanned Aerial Vehicles (UAV). Examples of supporting systems are obstacle warning systems in cars, and autonomous landing for aircraft. Also various projects and researches are related to UAV and they are utilizing in many fields. It is getting usable at almost everywhere, from military scouting to civilian production facilities.

In this paper, we propose a vision-based hovering control system for UAVs to maintain stable attitude by real-time image processing. A UAV equips many of sensors to measure inertial, altitude, acceleration and position of vehicle. Furthermore, most of them install image sensor for high-level missions like tracking target or transmitting a sequence of images. We make an approach to grasp a situation of vehicle by image information. There are a lot of horizontal and vertical lines in real world; we concentrate them as indicators of vehicle's attitude. Line is fundamental feature to comprehend circumstances. By detecting and tracking of lines, we can judge the status of vehicle.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_52,

2 Related Works

Many of researches for intelligent UAVs are related to computer vision and image processing. Line tracking, as one of the low-level methods to extract useful features from the image, is very useful technique that can be used in some practical implementations. There are a few researches dedicated to line detection and on the image. Most of these researches take a Hough transform technique as a core part of algorithm. Some of studies have been done in order to solve the computational problem [1][2].

Compute Unified Device Architecture (CUDA) is parallel computing platform and programming model invented by NVIDIA and implemented by the graphic processing units (GPUs) [7][8]. GPUs can achieve a high computing performance due to many of processing units and different memory spaces.

CUDA gives researchers access to the virtual instruction set and memory of the parallel computational elements in CUDA GPUs. Unlike CPUs, however, GPUs have a parallel throughput architecture that emphasizes executing many concurrent threads slowly, rather than executing a single thread very quickly.

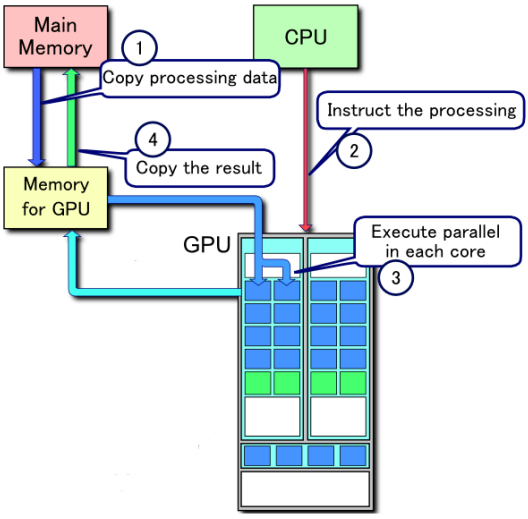


Fig. 1. Processing flow on CUDA

CUDA has several advantages over traditional general-purpose computation on GPUs (GPGPU) using graphics APIs. One of them is code can read from arbitrary address in memory. Second is it exposes a fast shared memory region and it can be used as a user-managed cache. Fig. 1 shows a brief processing flow on CUDA architecture.

In CUDA architecture, the first step is copying data from main memory to GPU memory. Secondly, CPU instructs the process to GPU. Thirdly, GPUs execute parallel in each core. Finally, it copies the result form GPU memory to main memory.

In this paper, we deal with CUDA via OpenCV (Open Computer Vision) software library [6]. OpenCV includes GPU module that contains all GPU accelerated stuff.

3 Implementation of Line Detection and Tracking on CUDA

3.1 Detection of Lines

In this section, detection algorithm which is based on unsupervised clustering is explained.

Detection algorithm is similar to tracking algorithm with some modifications. At the first frame specified for detection preprocessed and enhanced. Step3 searches for edges on preprocessed image and select probable positions of line segments. A binary image obtained on the previous step then used for calculating a full Hough space, in order to detect any line in whole position. Result of Hough space usually contains a few peaks that show strong lines on image. Then, detection algorithm thresholds a Hough space with some predefined value with a subsequent clustering to expose these lines. On the last step, algorithm searches local maximums around cluster centers to make sure that the line is aligned to the best suitable position on the image.

Main steps of algorithm are composed as follows.

- Step 1: Receive a new frame
- Step 2: Apply a preprocessing procedures
- Step 3: Edge detection
- Step 4: Convert image to Hough Space (full conversion)
- Step 5: Threshold the Hough space by some value
- Step 6: K-means clustering [3] with Mahalanobis distance metric
- Step 7: Align a lines positions

3.2 Tracking Lines

This section contains a description of line tracking algorithm. Proposed tracking algorithm in this paper is based on two main techniques. Firstly, Hough Transform [4] is used to transfer image to the Hough space, where we can identify some primitive shapes like line and circle. Secondly, a set of Kalman filters [5] are used to predict line's position in next frame to reduce the searching area and consequently improve overall performance of algorithm. Experiments show that these techniques work well in combination. Tracking steps are same to detection ones from 1 to 3. From step 4 to step 6 listed below.

- Step 4: Convert image to Hough Space (partial conversion)
- Step 5: Search most probable lines around predicted positions
- Step 6: Update lines positions and predict next position using sets of Kalman filter

3.3 Implementation on CUDA

Proposed algorithm needs to calculate probabilities of all possible lines around previous position in some region in case of tracking or all possible lines in case of detection. Simple calculation shows, that filling accumulator up procedure requires a certain number of operations with trigonometric functions.

Obviously, even modern computers will be unable to execute so many operations fast enough to make it in real time. To overcome this problem a few treatments has been done to improve the performance of algorithm. One is optimization on code level and the other is utilizing a GPU.

Fortunately, the voting process that fills accumulator in Hough space can be easily separated to parallel sub-tasks. And each sub-task can be processed independently; therefore, Hough transform algorithm part fits to CUDA architecture. In our implementation, all voting process in Hough space fully implemented on GPU, that provides huge bust in speed-up, also this acceleration increases with rise of input data resolution. The function, calculating the Hough accumulator, is fully executed on GPU.

4 Experimental Results

4.1 Experimental Environment

To make an implementation for our research, OpenCV (Open Source Computer Vision) [6] is used. OpenCV is a library of programming functions for the real-time computer vision. The library has more than 2500 optimized algorithms in image processing and machine vision areas. Some of these algorithms are also implemented on CUDA. The image processing is operated on the general purpose x86-machine with a GPU on the board. We made a measurement with graphical processing unit: NVIDIA Quadro 2000D. CUDA SDK 5.0 is installed to develop algorithm utilizing CUDA features.

4.2 Performance Analysis

Experimental result provides a comparison between full Hough space calculation, used for detection algorithm, on CPU and GPU. The resolution of Hough space is 360 degrees with frame resolution 320x240, and there are 144,000 Hough points need to calculate. Fig.2 shows comparison of Hough space calculation between CPU and GPU. Upper line means elapsed processing time for a frame on CPU at detection, and bottom line means time on GPU. Utilizing GPU provides average speed up around 4.5 times per second in comparison with CPU implementation. In terms of frames per second it equals 5 fps on CPU and 17 fps on GPU. This result shows that using GPU makes the possibility to detect lines in real-time.

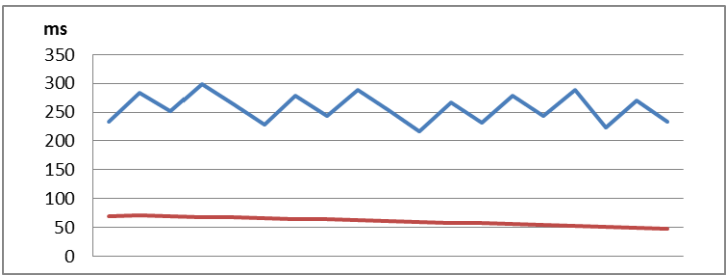


Fig. 2. Comparison of Hough space calculation on CPU and GPU

5 Conclusion and Future Work

In this paper, we present image processing system for vision-based hovering control which is implemented on CUDA. Some of critical requirements for control UAV by vision information are satisfying frame rate and latency in real-time. Our approach shows it is feasible to apply on real system. Implementations of detection and tracking on GPU work fast sufficient to apply UAV system.

At this moment in time, portable device does not have powerful performance enough to perform image processing, however, there will release in near future. Parallel computing is becoming common scheme in embedded system for some applications. In the next paper, we will show the result with real flight.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number: 2012006817).

References

1. Mills, S., Pridmore, T., Hills, M.: Tracking in a Hough Space with the Extended Kalman Filter
2. Choi, K., Min, K., Lee, S., Park, W., Seo, Y., Hong, Y.: Lane Tracking in Hough Space Using Kalman filter
3. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7) (July 2002)
4. R.O. Duda, P.E. Hart: Use of Hough Transformation to detect lines and curves in pictures (April 4, 2003)
5. Welch, G., Bishop, G.: An Introduction to the Kalman Filter, Department of Computer Science, University of North Carolina (2006)
6. `http://opencv.org`
7. Chen, S., Jiang, H.: Accelerating the Hough Transform with CUDA on Graphics Processing Units, Department of Computer Science, Arkansas State University
8. Gómez-Luna, J., González-Linares, J.M., Benavides, J.I., Zapata, E.L., Guil, N.: Parallelization of the Generalized Hough Transform on GPU, Computer Architecture and Electronics Department, University of Córdoba, Córdoba, Spain

Finding Relationships between Human Affects and Colors Using SVD and pLSA

Umid Akhmedjanov[1], Eunjeong Ko[1], Yunhee Shin[2], and Eun Yi Kim[1,*]

[1] Visual Information Processing Lab.,
Konkuk University, Seoul, Korea
{aurshax89,goejeong85,eykim04}@gmail.com
[2] Korea Intellectual Property Service Center, Seoul, South Korea
yhshin@kipa.org

Abstract. In this paper, a new method is presented to automatically find relationships between human affects and colors. For this, the probabilistic latent semantic model analysis (pLSA) and singular value decomposition (SVD) is applied. The proposed method is composed of three modules: feature extraction, feature transform and pLSA training. We first segment the image using mean-shift clustering, then extract color compositions by analyzing the colors from one region and its adjacent regions. Next, for the occurrence matrix, the SVD and pLSA are used. Using SVD, the occurrence matrix is decomposed into rank and null space matrix, where the null space is discarded and only the space corresponding to the singular values is used for further processing. For the reconstructed matrix, the pLSA is applied to obtain the correlation between affective classes and color compositions. To assess the effectiveness of the proposed system, it was applied to index the images using human affects. Then the results showed the effectiveness of the proposed method.

Keywords: Affective Mapping**,** Probabilistic Affective Model, Singular Value Decomposition, Probabilistic latent semantic analysis.

1 Introduction

With increasing the importance of affective computing, it becomes necessary to retrieve and process images according to human affects or preferences [1-3]. Among several image domains like photographic, medical or art, especially, it is very important to use affects on photometric image retrieval [1, 3].

Fig. 1 shows examples of searching 'mountain' pictures according to affects, where they are manually annotated by users. Those images have different feelings, despite of the images belong to the same object class. Thus, to distinguish them, it is important to predict affects from image.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_53,

Gorgeous Natural

Fig. 1. Search example of 'mountain pictures in accordance with affective classes

So far many algorithms and systems have been developed to index images using affects and objects. Most of them use the machine learning technique to find the latent semantics such support vector machine (SVM) and latent semantic model (LSA). Datta *et al*. investigated the relationship between several visual features and the aesthetic quality [3]. This method was used to evaluate aesthetic on painting and photographic images. However, in practice, there are many affective features useful for image retrieval, such as 'romantic' and 'gorgeous'. Therefore we need to consider using more affective features and to predict them. Kobayashi *et al*. considered more various affects than others [4]. Kobayashi's affective classes are composed of 180 adjective words which were subdivided into 15 groups. Thereafter, they have built the relation between 180 affective classes and colors, based on market research for several years, which is time consuming.

Accordingly, this paper presents a new method to automatically find relationships between several human affects and colors.

2 Proposed System

The goal of the proposed system is to find relationships between human affects and colors. For this, the proposed system is performed by three steps: feature extraction, feature transform and pLSA training, as shown in Fig. 2. We first segment the image using mean-shift clustering, then extract color compositions by analyzing the colors from one region and its adjacent regions. Next, for the occurrence matrix, the SVD and pLSA are used. Using SVD, the occurrence matrix is decomposed into rank and null space matrix, where the null space is discarded and only the space corresponding to the singular values is used for further processing. For the reconstructed matrix, the pLSA is applied to obtain the correlation between affective classes and color compositions.

Fig. 2. An overview of the proposed system

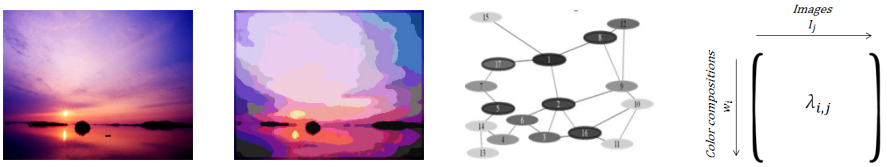
2.1 Feature Extraction

Generally, an image is depicted by some visual features, such as color, texture, shape and so on. Among them, color composition is only considered in current work. To

extract color compositions from a given image, this module is performed by two steps: image segmentation and color composition extraction.

In image segmentation, the image is partitioned into clusters using similar colors. We use the mean-shift clustering algorithm to divided an input image into regions, $R = \{R_i | i \in N\}$. As the increasing of number of regions N, considering those all combinations is computationally intensive. Therefore, we extract the color compositions for more important regions rather than all regions, which are called as seed region. To find seed regions, the importance value of region is calculated by multiplication of $A(R_j)$ and $G(R_j)$, where $A(R_j)$ is area of the current region and $G(R_j)$ is the Gaussian distance from its center to the centroid of image. Based on the importance values, we rank the regions, and select the top M regions as seed regions, $S = \{S_1, S_2, \ldots, S_M\}, (M < N)$. From only the seed regions, the color compositions between itself and its two adjacent regions are extracted and stored onto $\Lambda = \{\lambda_i\}$.

Figure 3 shows the results of feature extraction module. To generate the occurrence matrix, we used 500 training images, which were downloaded from Google using 10 queries.



(a) (b) (c) (d)

Fig. 3. Example of feature extraction results: (a) input image, (b) segmented result, (c) region adjacency graph and (d) generated occurrence matrix

2.2 Feature Transform

To handle the space sparsity, we use the SVD to reduce the null space and to transform the more meaningful space. Currently, the occurrence matrix X is decomposed into the product of three other matrices as following equation (1).

$$SVD(X) = U\,S\,V^T, \tag{1}$$

where U and V are orthogonal matrices and S is a diagonal matrix. S contains the square roots of the singular values in decreasing order. Then, approximately zero values of S are considered as 'noise' and discarded. In order to reconstruct the matrix, the corresponding row vectors of U and the corresponding column of V^T are also eliminated.

In the proposed system, the occurrence matrix X is generated from 500 images, which is represented by 1,170 color compositions. Thus, the matrix sized at 500 x 1170, and it is transformed to more efficient space in find the correlation between human affects and colors. Then, most elements of the matrix X are close to zero value. So, these values make difficult to discriminate the images based on human affects. By adjusting eigenvalue, we transformed the occurrence matrix and set to 100 as the optimal value.

2.3 pLSA Training

To automatically predict the affective features from transformed occurrence matrix, the pLSA is used as learning algorithm to find hidden human affects from colors. A scale represents correlation between affects and colors. The key concept of pLSA is to map the high-dimensional occurrence vector to a lower dimensional affective vector, i.e. generate the scale. Co-occurrences probability $P(I,W)$ is shown as following Equation (2).

$$\boldsymbol{P(I,w) = P(I)\sum_{z\epsilon Z} P(I|z)P(w|z)} \tag{2}$$

Here, *I* and *W* mean given image and extracted color compositions from the image, respectively. *Z* is the affective vector. Then, $P(I|z)$ is ground truth and $P(z|w)$ is scale, which represents correlation between affects and color feature. This probability is approximated by Expectation Maximization (EM) algorithm. The equation (2) can be rewritten like equation (3). Using EM, two probabilities are updated by following equation (4) and (5).

$$\boldsymbol{P(z|I,w) \propto \frac{P(z|I)P(w|z)}{\sum_{z\in Z} P(z|I)P(w|z)}} \tag{3}$$

$$\boldsymbol{P(z|I) \propto \sum_{i\in I} n(I,w)P(z|I,w)} \tag{4}$$

$$\boldsymbol{P(w|z) \propto \sum_{w\in W} n(I,w)P(z|I,w)} \tag{5}$$

3 Experimental Results

This study developed system that can find correlation between the human affects and visual features using SVD and pLSA. To assess the effectiveness of the proposed system, we implemented affects-based image indexing system using scale trained by the proposed system. We have conducted efficiency test using 1,485 web images and compared its performance with one of the existing system [1].

3.1 Affects-Based Image Indexing System

We developed probabilistic affective model (PAM) for predicting some affective classes that are associated with a photographic image from color features. The goal of PAM is to transform a given image *i* onto affective vector *E(i)* shown in following equation (6).

$$E(i) = (e_{i,z_1}, e_{i,z_2}, \ldots, e_{i,z_L}) \tag{6}$$

$$e_{i,z_m} = p(z_m|i) = p(w|i)p(w|z_m) = \prod_{j=i}^{N}\prod_{p=1}^{Pj} p\left(w_{j,p}|i\right)p(w_{j,p}|z_m) \tag{7}$$

In equation (7), affective feature e_{i,z_m} means the probability that the image i belongs to the corresponding m^{th} affective class ($w_m \in W$). In order to calculate the affective vector, visual features first are extracted from image and those features are mapped into affects using scale. $p(w_{j,p}|i)$ indicates the probability to map color compositions from image i and $p(w_{j,p}|z_m)$ is called as *scale* trained by the proposed system. The affective vector *E(i)* is calculated by multiplying scale and the conditional probability for obtained color composition from image *i*.

In this paper, six affective classes are used as followings: *pretty*, *colorful*, *dynamic*, *wild*, *romantic*, *pure*. Using PAM, color composition vector obtained from an image is mapped into six affective classes.

3.2 Results

To evaluate performance of the proposed system, the recall and precision are used. The results were compared with one of the existing system [1]. Figure 4 shows the results. As can be seen, the proposed method can be almost superior to the existing system.

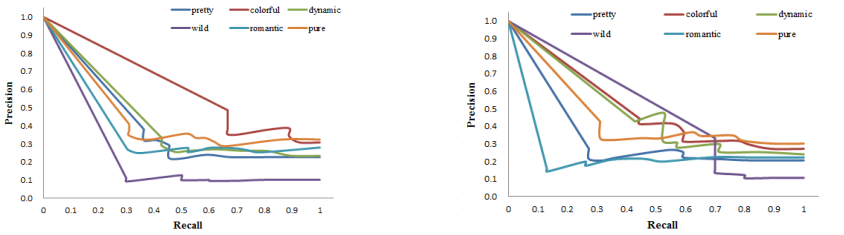


Fig. 4. Precision-recall curves of previous system (left) and the proposed system (right)

Acknowledgements. This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA(National IT Industry Promotion Agency.

References

1. Shin, Y., Kim, E.Y.: Affective Prediction in Photographic Images Using Probabilistic Affective Model. In: CIVR 2010 (2010)
2. Yashar, M., Benjamin, P., Joemon, J.: Handling Data Sparsity in Collaborative Filtering using Emotion and Semantic Based Features. In: SIG IR 2011 (2011)
3. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part III. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
4. Kobayashi, S.: Color Image Scale. Publishing of Kodansha (1991)

Home Appliance Control and Monitoring System Model Based on Cloud Computing Technology

Yun Cui[1], Myoungjin Kim[1], Seung-woo Kum[3], Jong-jin Jung[3], Tae-Beom Lim[3], Hanku Lee[2,*], and Okkyung Choi[2]

[1] Department of Internet & Multimedia Engineering, Konkuk University
[2] Center for Social Media Cloud Computing, Konkuk University, Neungdong-Ro 120, Gwangjin-Gu, Seoul, Republic of Korea
[3] 3 Digital Media Research Center, Korea Electronic Technology Institute, Electronics Center #1599, Sangam-dong, Seoul 121-835, Republic of Korea
{ilycy,tough105,okchoi20,hlee}@konkuk.ac.kr,
{swkum,mozzalt,tblim}@keti.re.kr

Abstract. With the development of intelligent home appliance technology, real-time home appliance status information is now generated in large quantities. New technology is necessary in order to process the large amount of status information that is generated every day. An innovative technology that has recently been used to process large amounts of data is cloud computing. Therefore, in this paper, we propose a system model to control and monitor home appliances using home network and cloud computing technologies in a smart home environment. UPnP technology is used to extract status information from home appliances. Cloud computing technology analyzes and processes the information and also provides virtualization services to users. In the proposed method, the gateway collects and stores home appliance information using home network technologies and sends the information to the cloud server for storage and management.

Keywords: cloud computing technology, UPnP, virtualization services, smart home.

1 Introduction

Users are constantly provided with more convenient services due to the development of a variety of computing techniques such as mobile communication technology and data processing techniques. As the development of home appliance control techniques, embedded techniques, and smart devices has progressed, innovative smart home technology has also developed in the last several years. However, it is difficult to store, process, and manage the large amounts of status information that is generated by the home appliances of a smart home on a single node (i.e., PC). A collection of regional smart homes generates a huge amount of data per day. Therefore, a

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_54,

technology to store and process a large amount of data is urgently needed. The technology that can provide such a technique is cloud computing. Cloud computing is a very efficient technique to process and analyze large amounts of data. Consequently, in this paper, we propose a novel system to collect the data generated in smart homes and process it based on cloud computing technologies.

The proposed system model is divided into three parts: the gateway, cloud server, and smart device. The gateway identifies home appliances that use UPnP services, extracts the status information of the home appliances, and transmits the extracted data to a cloud server [2], [3], [8]. The cloud server stores data classified by the user and provides home appliance monitoring services to users using the virtualized status information of the home appliances. It also offers a distributed computing function and data storage service to users via Hadoop-based technologies such as MapReduce, HDFS, and Hbase [1], [3], [10]. The smart device allows users to monitor and control home appliance functions. The smart device receives the virtualized data of the home appliances from the cloud server. All of the proposed components communicate with each other using HTTP and transmit data using XML [8].

The remainder of this paper is organized as follows: Section 2 discusses current research work related to smart home technologies. Section 3 describes the proposed system architecture and explains its main functions. Finally, we conclude the paper in Section 4.

2 Related Work

Cloud computing technology provides functions to store, handle, and manage large amounts of data over the internet. It consists of three platforms, IaaS, PaaS, and SaaS to offer many kinds of services to users.

IaaS (Infrastructure as a Service) is a platform as well as the base layer of the cloud computing environment offering computing resources for cloud computing. The core technology of the IaaS platform is virtualization, a technology that converts physical resources to virtual resources and increases resource utilization and flexibility. PaaS (Platform as a Service) is a platform that uses high-speed distributed computing technology to process the large numbers of media data requests from SaaS and manages the data storage in IaaS. SaaS (Software as a Service) is a service platform to provide data management APIs and web-based development tools in the cloud computing environment [3].

A smart home, in the conventional sense, supports automatic systems to control lighting and temperature and activate security apparatus. It is used to monitor many aspects of daily life [4]. Nowadays, smart homes incorporate many computing technologies to provide convenient personalized service to users within the home network [5]. Recently, much research on the smart home has focused on the home gateway. Using a home gateway, a smart home can form a peer-to-peer network to provide home network service anytime, anywhere [6]. The lack of a de facto communication standard for smart homes hinders the integration of different services. Therefore, Kim et al. proposed smart home software architecture based on OSGi (Open Services Gateway initiative) [7].

We reference these concepts when designing a system model to provide a personalized smart home service using cloud computing technologies.

3 Proposed System Model

The proposed system provides home appliance monitoring service to users via a home appliance virtualization function supported by a cloud server. The cloud server also stores the status information of the home appliances transmitted from the gateway of the smart home and uses cloud computing technology to process this information. The gateway consists of a PC and home network using UPnP to search and collect the metadata of the home appliances in a smart home. The collected metadata is transmitted to the cloud server by the gateway. Figure 1 shows the architecture of the proposed system model.

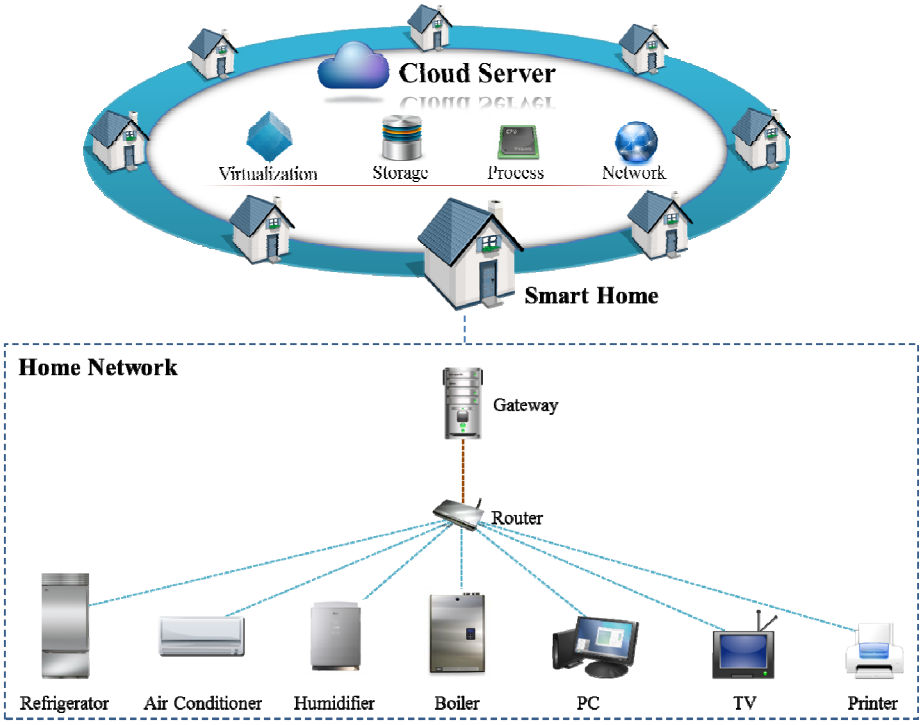


Fig. 1. Proposed system model architecture diagram

The gateway is an important component of the smart home because it acts as a network bridge between the cloud server and the smart home. The gateway also monitors the home appliance statuses, collecting home appliance status information every ten minutes. If a home appliance status changes, the gateway gets this information and sends it to the cloud server. The gateway is composed of four modules to provide home appliance control functions as follows. The identification

module authorizes separate users using OAuth [9] generated by the cloud server. The UPnP device control module provides device selection, registration, and status monitoring functions using SSDP and GENA. The device controlling module sends actions to control the home appliances. The device metadata extraction module extracts detailed metadata from the home appliances. The device log data transmission module sends the data to the cloud server.

The cloud server manages data, but remains separate from a smart home. Thus, the cloud server stores and manages home appliances status in a smart home over the gateway. The cloud server distinguishes and stores each home appliance of every smart home by user ID. The user ID is managed by the cloud server using Hbase, generating an OAuth-based identification number for each user. The cloud server virtualizes the home appliance status data of each independent smart home to provide virtualization service to users by user ID. The cloud server sends the virtualized home appliance data to the user's smart devices to support home appliance virtualization services.

The user's smart device communicates with the cloud server and receives the status information of the home appliances using a specific application. The smart device includes an XML data extraction module. This module is able to extract XML data transmitted from the cloud server. All of the components communicate with each other using XML data transmitted through HTTP. If users are at home with a smart device that communicates with the cloud server, the smart device is able to directly connect with the gateway over the home router, reducing unnecessary traffic among the gateway, cloud server, and smart devices.

4 Conclusion

In this paper, we propose a cloud-based system model to provide real-time home appliance monitoring and control services. The cloud server stores and manages a large amount of the status data generated by home appliances in smart homes. For transmitting cloud server status data of home appliances, we designed a gateway to support data communication between the cloud server and smart homes. The status data of home appliances in the cloud server is transferred to a smart device. Users monitor and control home appliances through the smart devices using the virtualized status data from the cloud server. However, if many users connect to the cloud server at the same time, a high amount of network traffic will occur in the cloud server. Therefore, to support better load dispersion, we will research effective algorithms in our future work.

Acknowledgments. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the C-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Dean, J., Ghemawat, S.: Mapreduce: Simplified Data Processing on Large Clusters. In: 6th Symposium on Operating Systems Design and Implementation, OSDI 2004, pp. 137–150 (2004)
2. Dugeon, O., Mahdi, M., Bars, R., Carbou, R.: Extended UPnP Multimedia Content Delivery with an HTTP Proxy. In: Zeadally, S., Cerqueira, E., Curado, M., Leszczuk, M. (eds.) FMN 2010. LNCS, vol. 6157, pp. 87–99. Springer, Heidelberg (2010)
3. Cui, Y., Kim, M., Lee, H.: Social Media Sharing System: Supporting Personalized Social Media Service Using UPnP Technology in Cloud Computing Environment. Information 15, 2043–2054 (2012)
4. Bregman, D.: Smart Home Intelligence – The eHome that Learns. International Journal of Smart Home 4(4), 35–46 (2010)
5. Lin, H.T.: The Development of Control and Energy Usage Information Modules for Smart Homes. In: ICCAIS 2012, pp. 236–240 (2012)
6. Wei, Z., Qin, S., Jia, D., Yang, Y.: Research and Design of Cloud Architecture for Smart Home. In: ICSESS 2010, pp. 86–89 (2010)
7. Kim, J.E., Boulos, G., Yackovich, J., Barth, T., Beckel, C., Mosse, D.: Seamless Integration of Heterogeneous Devices and Access Control in Smart Homes. In: 2012 8th International Conference on Intelligent Environments, pp. 206–213 (2012)
8. Contributing members of UPnP Forum.: UPnP Device Architecture Version 1.1. In: UPnP Forum (2008)
9. Hammer-Lahav., E.: The OAuth 1.0 Protocol. Internet Engineering Task Force, RFC 5849 (2010)
10. Vora, M.N.: Hadoop-HBase for Large-Scale Data. In: 2011 International Conference on Computer Science and Network Technology, pp. 601–605 (2011)

Load Distribution Method for Ensuring QoS of Social Media Streaming Services in Cloud Environment

Seung Ho Han, Myoungjin Kim, Yun Cui, SeungHyun Seo,
Yi Gu, and Hanku Lee*

Department of Internet & Multimedia Engineering, Konkuk University,
Neungdong-Ro 120, Gwangjin-Gu, Seoul, Republic of Korea
{shhan87,tough105,ilycy,dekiller2,guyi1987,hlee}@konkuk.ac.kr

Abstract. As various types of smart devices have recently appeared, SNS (Social Networking Services) have been expanded. Thus, the demand for social media streaming is on the rise. In the previous study, a media conversion system for ensuring QoS (Quality of service) of media streaming was presented. The presented system implemented a distributed streaming environment with multiple servers in order to perform reliable streaming of converted media. The method of distributing streaming job is crucial in implementing a distributed environment. Thus, the presented system established distributed streaming servers that applied RR (Round Robin) and LC (Least Connection) algorithms. However, since systems that applied RR and LC do not consider CPU utilization rate and network transmission traffic, they have limitations on reducing the burdens of servers. This study will present a SRC (Streaming Resource-based Connection) scheduling algorithm for ensuring QoS in the distributed streaming environment. The focus of this SRC algorithm considering CPU utilization rate and transmission traffic of servers is resolving the limitations of existing algorithms. As a performance evaluation, utilization rate of different systems that each applied SRC, RR and LC will be compared.

Keywords: cloud computing, Social Media Streaming, Load Distribution, QoS.

1 Introduction

As various types of high-performance smart devices such as smartphone, smart pad, and smart TV have recently appeared and SNS (Social Networking Services) such as Facebook and Twitter have been expanded, the demand for social media streaming services for diverse environments (e.g. wired and wireless) offered anytime, anywhere is rapidly rising [1], [2]. However, since media streaming data has high network bandwidth, studies on ensuring QoS (Quality of Service) are necessary [3]. Such studies require technology that converts media data into a form suitable for streaming. However, existing internet computing bases are burdensome, because they require substantial resources for lots of changes in multimedia [7]. Therefore, the previous

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_55,

study presented a Hadoop-based multimedia conversion system using a distributed environment [4], [5]. The presented system is capable of performing distributed parallel processing of large amounts of massive media files generated from SNS fast in the form appropriate for streaming of various smart devices [4], [5]. For reliable streaming of converted media, a method of creating distributed environment with multiple servers is one option.

The method of allocating job to each server upon request from users in a distributed streaming environment is crucial. Thus, the presented system established distributed streaming servers that applied RR (Round Robin) and LC (Least Connection) algorithms. RR is a method of streaming by selecting streaming servers in a prescribed order upon video streaming request from users [8]. LC is a method of streaming by selecting the streaming server with the least number of user access upon video streaming request from users [8]. However, since systems that applied RR and LC do not consider CPU utilization rate and network transmission rate they have limitations on reducing the burdens of servers. Thus, this paper presents a SRC (Streaming Resource-based Connection) scheduling algorithm that considers CPU and streaming transmission rate usage of servers, with focus on resolving the limitations of RR and LC distribution methods. Since SRC algorithm considers CPU usage rate and transmission rate of servers, resolving the limitations of existing algorithms was the focus. A performance evaluation on the system that applied this algorithm will compare and analyze the network streaming transmission rate between SRC-applied system and systems that each applied RR and LC.

2 Background

Amid rapid increase in network rate following the development of internet technology, studies on load balancing through server distribution have been conducted [6], [9]. In particular, solutions based on distributed environments for web server, file server, and media streaming that require lots of network rate have been studied [7], [8]. The previous study presented the following system. As production and sharing of social media have revitalized with various smart devices and expansion of SNS, transcoding and streaming technology based on distributed environments have been studied in order to offer social media compatible with various types of smart devices. However, existing approaches were limited to parallel distributed processing by increasing the number of cluster machines. The method of processing transcoding by increasing the number of cluster machines has the drawback of incurring rising costs as the number of cluster machines increases and didn't consider the splitting and merging policy of media files necessary for transcoding. As a solution, Hadoop-based distributed video transcoding that applied MapReduce Framework was presented [4].

Multimedia streaming data requires high network bandwidth. Thus, solutions based on distributed environments using multiple streaming resource-based connection servers have been studied for reliable streaming [6].

The method of distributing streaming job is crucial in implementing a distributed environment. Thus, the presented system established distributed streaming servers that applied RR (Round Robin) and LC (Least Connection) algorithms.

3 Streaming Resource-Based Connection

The previous study presented a Hadoop-based multimedia conversion system using a distributed environment. In order to perform streaming of videos converted in the system presented, distributed streaming was implemented using multiple servers. To allocate streaming job for distributed streaming configuration, RR and LC methods were applied. However, algorithms of existing methods do not consider CPU usage rate and streaming transmission rate. Therefore, this study presented SRC with focus on resolving the limitations of existing algorithms. The following is 'SRC Pseudo Code' presented by this study.

```
1: While Server_Available // begin 'WHILE'
2: //define n= 1..n
3: IF Streaming_Request THAN
4:    FOR Streaming_Server 1 to Streaming_Server n THEN
5:          Streaming_Server_Usage n
6:             = CPU_Usage + Network Transmission rate
7:    END FOR
8: END IF
9: Streming_Response
10: Server_Usage_MIN_number
11:           =  NumberofMin(Streaming_Server_Usage n)
12:    IF Server_Usage_MIN_number == 1 THEN
13:        Streaming_Response
14:           =  Min(Streaming_Server_Usage n)_Server
15:    ELSE
16:        Server n
17:         =  MIN(Streaming_Server_Usage n of network
18:                   transmission rate)_Server
19:        Streaming_Response = Server n
21:     END IF
22: End Streaming_Response
23:END WHILE
```

Fig. 1. Pseudo code of SRC Algorithm

Figure 1 shows the Pseudo code of SRC Algorithm. The streaming system was implemented by applying the SRC algorithm suggested by this study. SRC algorithm distributes streaming job in the following order. First of all, a streaming request is made from the user. Second, SRC-applied management module identifies the CPU usage of each streaming server and creates a ranking of utilization rate. Finally,

streaming services are provided to users via server with the lowest utilization rate. If more than one server has the same ranking, streaming services are provided to users via server that uses the lowest network transmission rate.

4 Performance Evaluation

Performance evaluation was conducted on 10 units of HDFS (Hadoop Distributed File System)-based contents server, 3 units of NginX-based streaming server, and 1 unit of cluster server with web-based management server. For content servers, HDFS 1.0.4 version was used, and for streaming server configuration, NginX 1.2.7 version and H.264 streaming module 2.2.7 version were used. In addition, management servers for interface and streaming distribution for streaming were established with Jsp in Tomcat7 environment, and CPU usage rate and network transmission rate measurement and streaming performance testing tools were established with Java. As a data set for performance evaluation, 4MB MP4 file converted from the presented system was used, and the protocol used for streaming is HTTP Progressive Download. For performance evaluation, 600 users accessed the systems that each applied RR, LC, and SRC to calculate the overall transmission rate generated upon media streaming, and average transmission rate per second was measured using the streaming time.

Table 1 shows the results of performance measurement. It revealed that SRC method-based algorithm that considers CPU and network had higher efficiency than RR and LC algorithms that do not take into account computing resources.

Table 1. Comparison of performance evaluation

	RR	LC	SRC
Server A	62.85 MB / sec	62.34 MB / sec	67.06 MB / sec
Server B	63.75 MB / sec	62.81 MB / sec	66.99 MB / sec
Server C	62.76 Mb / sec	62.79 MB / sec	66.39 Mb / sec

5 Conclusion and Future Work

This paper has confirmed that streaming distribution methods of existing LC and RR methods are inefficient, and presents a SRC-based distribution method as a solution, along with performance evaluation. Based on the results, improvement methods of SRC will be studied by considering the evaluation factors of the overall system such as DISK I/O and OS, as an extension of current study. Such research will contribute to ensuring the QoS of increasing social media streaming services.

Acknowledgments. This work was supported by the Konkuk University.

References

1. McAfee, A., Brynjolfsson, E.: Big data: the management revolution (2012)
2. Cisco Visual Networking Index: forecast and Methodology 2011-2016 (2012)
3. Ma, K.J., Bartoš, R., Bhatia, S.: A survey of schemes for Internet-based video delivery. Journal of Network and Computer Applications 34, 1572–1586 (2011)
4. Kim, M., Han, S., Cui, Y., Lee, H., Jeoung, C.: A Hadoop-based Multimedia Transcoding System for Processing Social Media in the PaaS Platform of SMCCSE. KSII Transaction on Internet and Information Systems (TIIS) 11, 2827–2848 (2012)
5. Heo, N., Lim, D., Seo, D., Jung, I., Kim, Y.: Load Distribution Method and Admission control for Streaming Media QoS in Distributed Transcoding Servers. In: ICCSA 2007, pp. 39–45 (2007)
6. Li, C., Peng, G., Gopalan, K., Chiueh, T.: Performance guarantee for cluster-based internet services. In: 2002 Ninth International Conference on Parallel and Distributed Systems, pp. 327–332 (2002)
7. Tep, Y.M., Ayani, R.: Comparison of load balancing strategies on cluster-based web servers. In: Simulation, pp. 185–95 (2001)
8. Piórkowski, A., Kempny, A., Hajduk, A., Strzelczyk, J.: Load Balancing for Heterogeneous Web Servers. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2010. CCIS, vol. 79, pp. 189–198. Springer, Heidelberg (2010)

A Robust Cloud-Based Service Architecture for Multimedia Streaming Using Hadoop

Myoungjin Kim[1], Seung Ho Han[1], Jong-jin Jung[3], Hanku Lee[1,2,*], and Okkyung Choi[2]

[1] Department of Internet and Multimedia Engineering, Konkuk University,
1 Hwayang-dong, Gwangjin-gu, Seoul 143-701, Republic of Korea
{touhg105,shhan87,hlee}@konkuk.ac.kr
[2] Center for Social Media Cloud Computing, Konkuk University,
1 Hwayang-dong, Gwangjin-gu, Seoul 143-701, Republic of Korea
hlee@konkuk.ac.kr, okwow2@gmail.com
[3] Digital Media Research Center, Korea Electronics Technology Institute,
Electronics Center #1599, Sangam-dong, Seoul 121-835, Republic of Korea
mozzalt@keti.re.kr

Abstract. Delivering scalable rich multimedia applications and services on the Internet requires sophisticated technologies for transcoding, distributing, and streaming content. Although cloud computing provides an infrastructure for such technologies, the specific challenges of task management, load balancing, and fault tolerance remain. To address these issues, we propose a cloud-based distributed multimedia streaming service, or CloudDMSS. The system is designed to run on all major cloud computing services, and is highly adapted to the structure and policies of Hadoop, which give it additional capabilities for transcoding, task distribution, load balancing, content replication and distribution.

Keywords: Streaming Service, Mobile Media Service, Cloud Computing, Media Transcoding.

1 Introduction

With the recent proliferation of rich social media across a variety of personal devices, considerable attention has shifted to the challenge of adaptively distributing and streaming multimedia content over the Internet. Among the technologies that have emerged, cloud-based media streaming, transcoding, and distributed storage have been the most noteworthy and influential.

The reason why cloud-based technologies have emerged in this regard can be discerned from the features of recent multimedia services: media heterogeneity, Quality of Service (QoS) heterogeneity, network heterogeneity, and device heterogeneity [1]. To support such features, the streaming, transcoding, and distribution of media must depend on massive—and massively scalable—computational resources, i.e., CPUs, memory, network bandwidth, and storage.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_56,

Though cloud computing can provide these resources, in doing so, it also introduce a heavy burden on existing Internet infrastructure and cloud resources, and it introduces a host of new challenges (e.g., cluster rebalancing, namespace management, data distribution/replication, auto-recovery, and fault tolerance), all of which are intensified under the massive swings in traffic associated with rich media streaming. These challenges have proven difficult for developers and service vendors alike, and continue to trouble current media delivery systems.

To address these challenges, we herein propose a cloud-based distributed multimedia streaming service (CloudDMSS) system designed to run on current cloud computing infrastructure. CloudDMSS capabilities include

(1) Transcoding of large amounts of media into the MPEG-4 video format for delivery to a variety of devices, including PCs, smart pads, and phones
(2) Exponential reduction in transcoding time through incorporation of the Hadoop file system (HDFS) for storage of multimedia data and MapReduce for distributed parallel processing
(3) Reduction in content delays and traffic bottlenecks using streaming job distribution algorithms
(4) Improvement in overall performance using dual-Hadoop clustering per physical cluster
(5) Efficient content distribution and improved scalability through adherence to Hadoop policies

The remainder of this paper is organized as follows: section 2 discusses relevant research on cloud-based streaming services; section 3 describes the core architecture of CloudDMSS with respect to transcoding, job distribution, content replication and distribution, etc.; section 4 presents our prototype of the proposed system and its configuration; and section 5 offers concluding remarks and plans for future work.

2 Related Work

In recent years, many researchers have applied cloud computing technologies to rich media services, in response to the explosion of demand for such services. This section presents the research most relevant to our CloudDMSS system.

Hui et al. in [5] proposed *MediaCloud*, a layered architecture that defines a new paradigm for dealing with multimedia applications and services. The architecture comprises three layers—a *Media Service Layer*, a *Media Overlay Layer*, and a *Resource Management Layer*—and addresses such key challenges as heterogeneity, scalability, and QoS provisioning. However, this architecture is treated mainly at the conceptual level, leaving most of the challenges of real-world implementation to future work [5]. In contrast, Luo et al. addressed the implementation challenge of QoS provision over virtualized infrastructure by presenting a practical architecture and mechanism for a private media cloud [3]. They describe their system in terms of four major components: monitoring, load balancing, traffic management, and security.

In regard to cloud-based streaming, Lee et al. in [4] have proposed a configuration scheme for connectivity-aware P2P networks with algorithms for connectivity-aware mobile P2P network configuration and connectivity-aware P2P network reconfiguration. In [6], Chang et al. described a cloud-based media streaming architecture that dynamically adjusts streaming services in response to mobile device resources, multimedia codec features, and network environment. They also presented a design for the stream dispatcher component, including real-time adaptation of codecs in response to client device profiling, and a dynamic adjustment of multimedia streaming (DAMS) algorithm. In [9], Huang et al. presented *CloudStream*, a cloud-based video proxy capable of delivering high-quality video streams by transcoding the original video in real time to a scalable codec, which in turn allows adaptation of the stream to various network dynamics. They also proposed a multi-level transcoding parallelization framework with two mapping options: hallsh-based mapping and lateness-first mapping.

3 Proposed System Architecture

Fig.1 shows an overview of the CloudDMSS architecture, highlighting three main modules: the Hadoop-based distributed multimedia transcoding module (HadoopDMT), and the Hadoop-based distributed multimedia streaming module (HadoopDMS), and the cloud multimedia management module (CMM).

The HadoopDMT transcodes a variety of multimedia data into MPEG4, a standard format in which media can be streamed and played on a variety of devices. Quality and speed are improved by adopting the Hadoop distributed file system (HDFS) [2] for storing video data from many sources, MapReduce [2] for distributed parallel processing of this data, and Xuggler [7] for transcoding the data.

Once transcoded, media contents are automatically moved and stored in HDFS of the HadoopDMS module. The contents are split into blocks of configurable size and distributed across the system. When a block is distributed, it is also replicated at three data nodes managed by NameNode, which constructs a directory tree of all transcoded contents according to the Hadoop distribution policy. This virtually guarantees content availability even under system or node failure. Furthermore, by conforming to the Hadoop policy, HadoopDMS automatically benefits from Hadoop's distributed processing capabilities, as well as its facilities for data replication, file splitting and merging, load balancing, and fault tolerance.

The role of the CMM module is to manage jobs such as streaming and transcoding tasks, and to balance the load on streaming servers with media stream scheduling. The module's streaming job distribution algorithm is called streaming resource-based connection (SRC). SRC optimally distributes streaming jobs among streaming servers in HadoopDMS based on CPU usage rates and the currently streaming traffic.

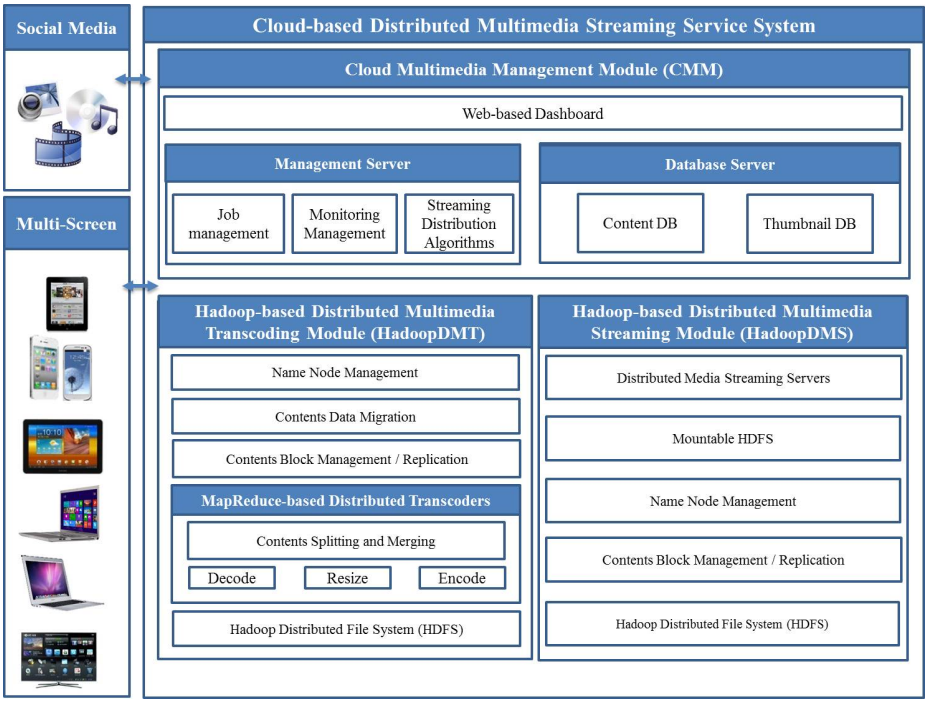
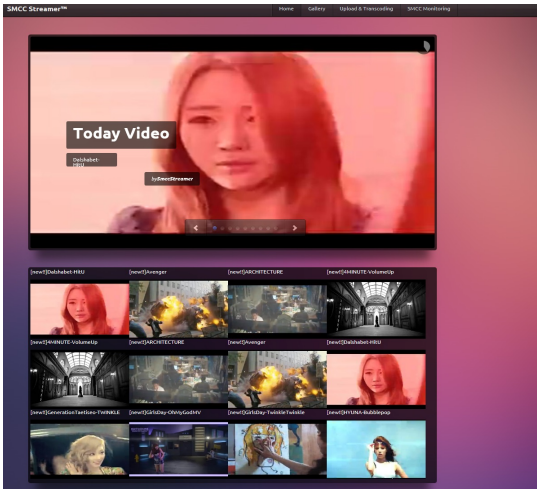


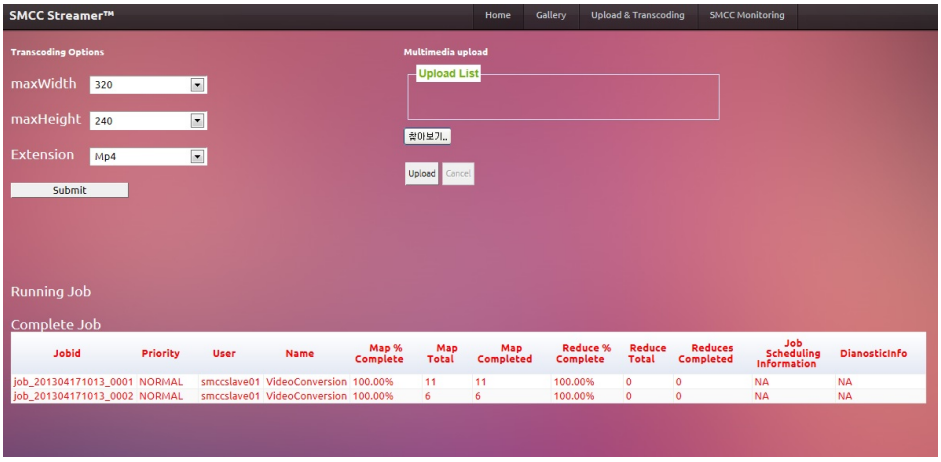
Fig. 1. Architectural overview of the CloudDMSS system

4 Implementation and Prototype

For our prototype implementation of CloudDMSS, we constructed our own cloud computing servers comprising 28 nodes in total. Each node consisted of Linux OS (Ubuntu 10.04 LTS) running on two Intel Xeon quad-core 2.13 GHz processors with 4 GB registered ECC DDR memory and 1 TB SATA-2 disk storage. All nodes were interconnected through 100 Mbps Ethernet adapters. To implement the CMM module, one node was designated as our management server, running Tomcat, and a second node was designated as the content DB server, running MySQL. The HadoopDMS module was composed of 1 NameNode and 12 DataNodes running on HDFS. The HadoopDMS module consisted of 3 streaming servers based on NginX and 10 content storage servers running on HDFS. We used a dual Hadoop cluster in one physical cluster to distribute load between transcoding and streaming tasks. Our software specification included Java 1.6.0_39 (64-bit), Hadoop-1.0.4, Xuggler 3.4.1012 (64-bit) for video transcoding, H.264 streaming module-2.2.7, and fuse_dfs_0.1.0.



(a)



(b)

Fig. 2. (a) Web-based dashboard for streaming transcoded contents (b) Web-based dashboard for transcoding tasks and resources

The output from the prototype system is provided in Fig. 2. Fig. 2(a) shows a selection of streamed content offered through a web-based dashboard, running commodity PC hardware. Fig. 2(b) shows a screenshot of the web page for managing transcoding tasks. Using this page, users and administrators can upload original content, select transcoding options (e.g., resolution, format, and codec) and stream the content to other users. Users can also monitor MapReduce-based transcoding processes and the remaining HDFS storage capacity.

5 Conclusion and Future Works

In this paper, we proposed the CloudDMSS system for efficient cloud-based streaming of rich social media. Our system addresses a number of pressing issues related to distributed media streaming, including transcoding for heterogeneous devices, job distribution, and content replication/distribution under HDFS.

Our current plan is to implement a fully functional CloudDMSS system, and to conduct thorough quantitative performance analysis of the system on a variety of cloud computing infrastructures, including Amazon EC2 and Rackspace Compute.

Acknowledgements. This research was supported by the MSIP (Ministry of Science, ICT & Future Planning) of Korea under the C-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Zue, W., Luo, C., Wang, J., Li, S.: Multimedia Cloud Computing. IEEE Signal Processing Magazine 28, 59–69 (2011)
2. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Communication of the ACM 51, 107–113 (2008)
3. Luo, H., Egbert, A., Stahlhut, T.: QoS Architecture for Cloud-based Media Computing. In: IEEE 3rd International Conference on Software Engineering and Service Science, pp. 769–772. IEEE press, Beijing (2012)
4. Lee, H.S., Lim, K.H., Kim, S.J.: A Configuration Scheme for Connectivity-aware Mobile P2P Networks for Efficient Mobile Cloud-based Video Streaming Services. Cluster Computing, 1–12 (2013)
5. Hui, W., Lin, C., Yang, Y.: MediaCloud: A New Paradigm of Multimedia Computing. KSII Transactions on Internet and Information Systems 6, 1153–1170 (2012)
6. Chang, S.Y., Lai, C.F., Huang, Y.M.: Dynamic Adjustable Multimedia Streaming Service Architecture over Cloud Computing. Computer Communications 35, 1798–1808 (2012)
7. Xuggler Java Library, `http://www.xuggle.com/xuggler/index`

Video Image Based Hyper Live Spatial Data Construction

Yongwon Cho[1], Muwook Pyeon[2,*], Daesung Kim[3], Sujung Moon[3], and Illwoong Jang[3]

[1] Department of Advanced Technology Fusion, Konkuk University, South Korea
avenuel88@gmail.com
[2] Department of Civil Engineering, Konkuk University, South Korea
neptune@konkuk.ac.kr
[3] Department of Advanced Technology Fusion, Konkuk University, South Korea
mutul94@snu.ac.kr, msujung@konku.ac.kr, jjang7551@konkuk.ac.kr

Abstract. Recently, Spatial information technology closer to reality three-dimensional space, a variety of information services and Web-based content, to provide information services through the space. CCTV video and camera video based multi-dimensional image data in order to build real-time spatial information-based and user-built CCTV video was up-loaded to the online video you should use them. In order to use the video upload large amounts of processing is required, spatial information can be presented as an alternative to building, large amount of data as a way to effectively use big data and cloud computing.

Keywords: Spatial Data, Hyper Live Map, service, Video Image, CCTV.

1 Introduction

Dimensional spatial information in the process of building a digital land acquisition refers to the various types of information, and this multi-dimensional spatial information, as well as non-spatial information services area of the region is required to deal with the information[2]. Recently, spatial information technology deployment and service to spatial information from the real-time three-dimensional space of the multi-dimensional space-based information services and Web-based content through a variety of services to users is changing to allow participation.[4]. Search real-time maps and spatial services various life events, and multi-dimensional space based on spatial information of the people and events of various ecological information provided by linking the concept of real-time multi-dimensional real-spatial information (Hyper-Live Spatial Data) to based techniques have been developed.

Also monitored by CCTV and video data as a possible, spatial information to provide in real-time demand for increasing services[1]. Naver Map Service and Daum Map Service in the normal map and satellite photo map services based on the offer to

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_57,

apply, and Naver in the street, aerial view, and Daum by providing a photo-realistic view of the load-based map services and are provided with real-time traffic information[1].

The services currently being provided to establish the cost of building a lot of expensive equipment, are built based on the image data, CCTV video and online spatial information services when building can be seen in terms of cost reduction. Also, upload your own videos and upload them built into the video if you use spatial information users can participate and every minute, every second of video are uploaded continuously updated, if the cost of the multi-dimensional real-time spatial information service that is expected to be available.

2 Image Data Acquisition Methods Research

2.1 CCTV

The purpose of the integration monitoring of local government and CCTV installation supplement is constantly increasing. CCTV has been used mainly for security purposes, in real time to provide via video as a service is required[1]. In the field of spatial information existing equipment instead of expensive LiDAR data designated place in order to reduce construction costs as fixed in real time utilizing the CCTV has been studied[5]. CCTV picture below is the video data using the Konkuk University main building.

Fig. 1. CCTV video data through the building

2.2 Camera Video

To mentioned above can be obtained through the CCTV image data, but only with CCTV images of any area cannot be obtained. So the image data in other ways to compensate people carrying smart phones and digital cameras, video data should take advantage of through user-participation. Smart phones and digital cameras very portable, and technology to the development of high quality images can be obtained easily[6], General with the consent of the user to upload images can be obtained through an internet search[7]. When these images are acquired to manage large amounts of images to solve this problem, which is currently active in the big data and cloud computing to take advantage of the one that you think is an alternative.

3 Fusion of Image Data to Suggestion

3.1 Big Data

Recently, GPS-equipped smart phone due to the prevalence of SNS activation, social media, due to the growth of diverse and numerous amounts of text, video, location data is generated in real-time position information, as well as the behavior of people and ideas and through SNS comments are able to analyze and predict. Most of this data collection, storage, retrieval, analysis, visualization, and difficult to non/semi-structured data, if not used properly useless, that can be data[10]. BigData and efficient processing of such data, analysis, and in order to take advantage of was the emergence, BigData is usually data volume, variety, velocity as a combination of three factors is characterized by changes[8]. Big Data and analysis techniques for processing such data, the text mining, opinion mining, social network analysis, cluster analysis has dual images similar to nested characteristics of the object together with the cluster analysis technique was used for outgoing[9]. CCTV video shown in Figure 1, starting with the image data, after defining the similarity of images in order from closest to the similarity overlap over the final I used for 3D modeling is built.

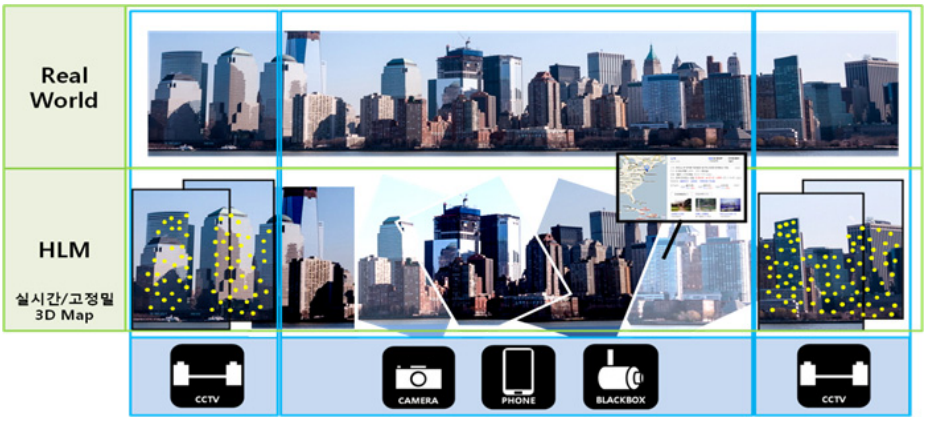


Fig. 2. Overlap process image

3.2 Cloud Computing

A necessary technology in handling big data cloud computing, the data throughput to other computers connected to the Internet, processing technology, and distributed processing over multiple servers is essential. The core technology of distributed processing, so the cloud is big data and close relationship.

The core technology of cloud computing, virtualization of the distributed processing[12]. What is virtualization server that actually processes the information the only that site is divided into multiple servers at the same time is a technology that enables multiple tasks You rate the utilization rate of the server can increase[11]. Distributed processing on multiple computers share processing tasks and collect the results back through the network is a way. The distributed system consists of a number of computer systems that behave as if they were one single computer system by large-scale operations can be processed quickly[11]. Thus, virtualization technologies and distributed processing technologies, if the data can have a greater efficiency in processing, so if you are using cloud computing technology to great effect can be obtained.

4 Build Multi-dimensional Real-Spatial Information

Image data representative of the services offered by the MS of the 'Photosynth'. 'PhotoSynth' in the case of individuals who hold multiple pictures, similar to the actual appearance by placing shall automatically complete three-dimensional picture. Three-dimensional pictures on-line store, and the location of the photos on the map by inserting has been able to show to other users.

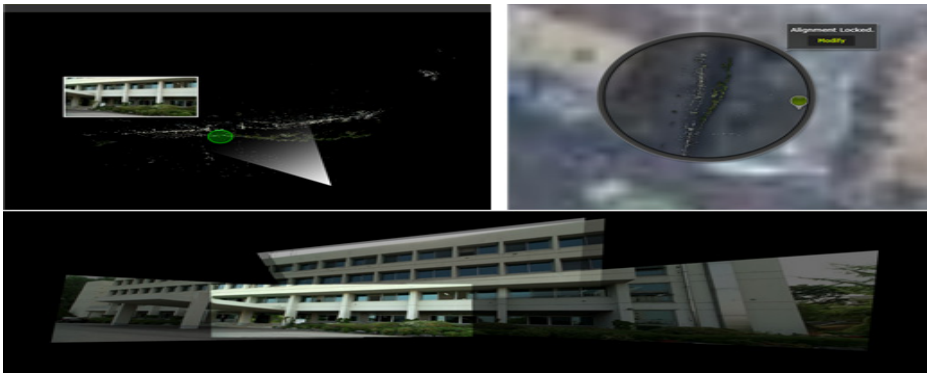


Fig. 3. 'Photosynth' 3D Point Cloud and inserted on the map (above), 3D photos (below)

Multi-dimensional real-spatial information (Hyper-Live Map) to in the image data obtained through the Point Cloud to 3D Reconstruction purpose of the study, performed experiments by CCTV.

Experimental results establish the structure only when recording CCTV, when shooting closed areas or shaded areas 50 percent due to the building was possible, To compensate for this, the video camera, or smart phone video by adding the results of building CCTV video was more effective than just that building. Figure 3 shows the Konkuk University main building with CCTV video camera image is using that constructed.

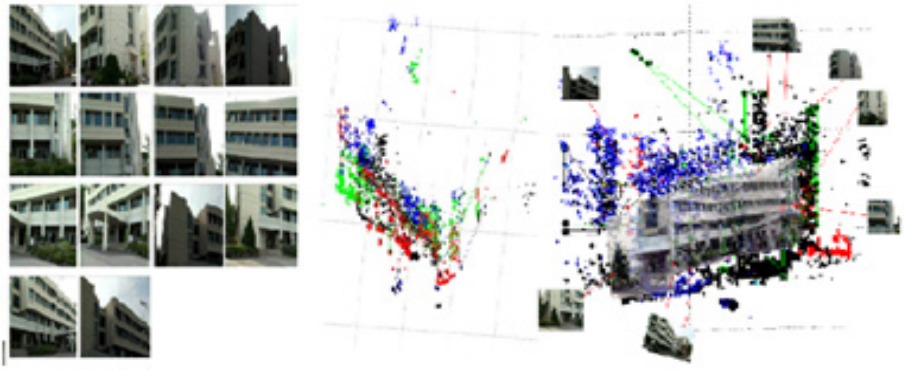


Fig. 4. Konkuk University main building CCTV video and camera video used to 3D reconstruction

In conclusion, if one unit of the structure of can be done above experiments, but with local scale multi-dimensional real-spatial information (Hyper-Live map), if services have a large amount of image data, and these data are classified with a high-resolution quality video the processing is necessary to separate. Build three-dimensional data generated while managing large amounts of data to solve the problem if they utilize big data and cloud computing for data processing speed increases, the effect of real-time spatial information will be exalted. And do not have expensive equipment utilization based CCTV video and online video built with just a big effect on the to get a building can bring about cost savings.

5 Future Research in Progress

This study CCTV or regular image data obtained through the camera of big data and cloud computing by leveraging real-time spatial information quickly and efficiently is possible that the service proposed. Upload your own image data and after up-loaded images that 3D data re-offer the advantage of being able to if the user will be able participation provide services. And the user's real-time spatial information due to the continuous video updates will be looks Service. The future matching the existing work that by image matching the data with the latest image data to build three-dimensional data, taking advantage of the big data and cloud computing that the experimental results proposed.

Acknowledgments. This research was supported by a grant from High-tech Urban Development Program - Development of Generation and Application Technology for Realtime Digital Map Project(11 High-Tech Urban G10) funded by Ministry of Land, Transport and Maritime Affairs & this work was financially supported by Ministry of Land, Transport and Maritime Affairs (MLTM) of Korea as part of 'U-City Master and Doctor Course Grant Program'.

References

1. Moon, S.J., Pyeon, M.W., Kim, C.J., Lee, S.W., Kang, N.G.: Element Analysis for Construction a Multi-dimensional Real-Time Map Service. In: ICONI (2012)
2. Lee, H.J.: Using multi-dimensional spatial information orthophoto production plan realized. Journal of Korea Society Surveying 26, 241–253 (2008)
3. Choi, S.K., Cho, E.H., Lee, B.Y.: Using multi-dimensional spatial information management facilities. Journal of Korea Society for Geospatial Information 18, 41–46 (2010)
4. Land, Transport and Maritime Knowledge and Information Centre, National Strategy Technology - Technology Trends in the major industrialized countries, Korea
5. Cho, Y.W., Pyeon, M.W., Kim, D.S.: Comparison of Image Matching Algorithms in the Spatial Information System. Journal of Korea Society for Geospatial Information (2013)
6. Meesters, L., IJsselstejin, W., Seuntiens, P.: Survey of perceptual quality issues in three-dimensional television system. In: Proc. of SPIE on Stereoscopic Displays and Virtual Reality System X, vol. 5006 (2003)
7. Salton, G.: Automatic Text Processing-the Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley Publishing Co., Reading (1989)
8. O'Reilly Radar Team, Planning for Big Data. O'Reilly (2012)
9. Kang, M.M., Kim, S.R., Park, S.M.: Analysis and utilization of Big Data. Journal of Information Science 30, 25–32 (2013)
10. Cho, S.W.: Big Data era technology, pp. 5–7. KT Institute of Technology, Korea (2011)
11. Armmrust, M.: A View of Cloud Computing. Communications of the ACM 53, 50–58
12. Vaquero, L.M., Merino, L.R., Caceres, J., Lindner, M.: A Break in the Clouds: Towards a Cloud Definition. In: ACM SIGCOMM Computer Communication, vol. 39, pp. 50–55 (2009)

A Peer-to-Peer Based Job Distribution Model Using Dynamic Network Structure Transformation

Seungha Lee[1], Yangwoo Kim[2], and Woongsup Kim[2,*]

[1] ETRI (Electronics and Telecommunications Research Institute), Daejeon, Korea
lesh915@etri.re.kr
[2] Department of Computer and Information Communications Engineering, Dongguk University, Seoul, Korea
{yangwoo,woongsup}@dongguk.edu

Abstract. Typically, many systems in organizations suffer from limited computer resources, while there are a huge number of under-utilized computers available which are able to contribute to continue reliable services when a system faces operational overloads. This paper proposes a Peer-to-Peer (P2P) based distributed job distribution model for job allocation and aggregation using Hub Peers. To this end, we first provide a peer-to-peer job distribution model using periodically collected peer information, and then proposed a peer structure transformation algorithm that composes P2P network topology dynamically for using under-utilized computing resources efficiently. Finally, we prove the benefits of our approach by comparing our proposed approach to other works.

Keywords: Peer-to-Peer system, distributed processing, job management, load sharing, tree transformation.

1 Introduction

Systems in organizations such as web portal sites often suffer from limited computer resources in handling the massive number of users. Such huge number of users triggers the system overloaded, and hence causes service unavailability and additional operational costs. Load distribution is a technique that is useful to such situation by distributing system-overloads to the under-utilized available computing resources, which are able to contribute to continue reliable services when a system faces operational overloads. Due to the advance of network technologies and the rapidly changing computing environment, there are many load distribution technologies such as NOW (Network of Workstation), Cluster, Grid [1,2], Peer-To-Peer (P2P) system [3], and Cloud computing [4]. P2P system is a two-way data exchange system between the independent network nodes (we call *peers*) through repeated interactions,

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_58,

and a distributed system that relies on distributed peers that may join and leave frequently.

In this paper, we propose a P2P based job distribution model for job assignment and aggregation using the Hub Peers. We propose a model to compose P2P based node topology dynamically such that jobs are efficiently distributed and aggregated based on a selected node called Hub Peer. To this end, we provide a Hub Peer selection algorithm and dynamic data exchange mechanisms using Hub Peers.

2 A Model for Load Distribution Model for P2P Environment

Our proposed P2P based load balancing model is based on structured P2P network using super peer which manage neighboring peer information. We call super peer as Hub Peer, as super peers in our approach do job distribution, aggregate job execution, route tasks as well manage peer information. In this section, we describe our Hub Peer selection algorithm, job distribution algorithm, and dynamic peer network structure transformation algorithm for P2P load distribution.

2.1 Hub Peer Selection Algorithm

For P2P based job distribution, a peer group, which collaborate each other, should be created and a Hub Peer managing the peer group must be elected. To create a peer group, our approach utilizes heartbeat messages, which are published to all the neighboring peers, exchanged among neighboring peers, used for counting the number of neighboring peers. The peer who gets most heartbeat message, that is the peer who has the most neighboring peers, is selected as Hub Peer and others remain as work peers. The Hub Peer must be located in close range to other peers in a peer group. So when a newly joining peer cannot find Hub Peer in close range, then a new peer group should be formed and elect new Hub Peer such that all peers in new peer group are located to the new Hub Peers in close range. In our Hub Peer selection algorithm, a newly joining peer sends peer selection message to one of neighboring peers. If the neighbor is already Hub Peer, then new peer can join the peer group immediately. Otherwise, the neighbor broadcasts heartbeat message to its surrounding peers in order to initiate new Hub Peer selection process. Each peer who get heartbeat message broadcast another heartbeat message to figure out how many neighbor peers are. The peer who has the most neighboring peer is selected new hug peer. (Fig. 1.)

2.2 Job Distribution Algorithm

A Hub Peer assigns jobs to its work peers and its neighbor Hub Peers, maintaining job assignment information. Hub peers has the information of maximum capacities and current workloads of its work peers and neighboring Hub Peers, and assigns new

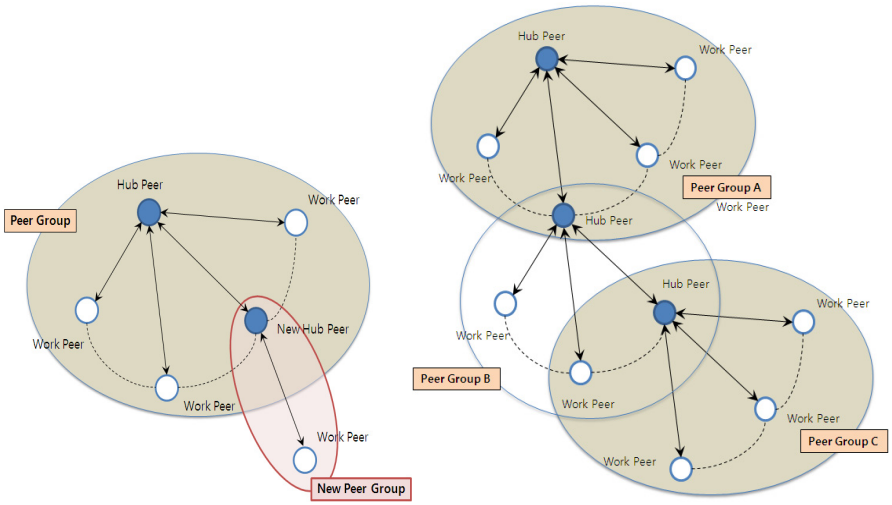


Fig. 1. A new peer joins to an existing peer group. New peer is not located beyond the existing Hub Peer's range, so new Hub Peer is elected (left). A new peer can join two existing peer groups A and C. As new peer is located beyond both A and C's Hub Peer range, new peer group B is formed by using heartbeat message (right).

tasks based on those information so that all peers should not have job assignment over their capacity. During job execution, every peer sends periodic message to Hub Peer, noticing the job is in execution or finished. If a job is still executing beyond the time threshold, then a Hub Peer considers the job execution fails and assign the job to another work peers. Each Hub peer i maintains a queue where total assigned jobs is recorded. Hub peers try to assign jobs to work peers or another Hub Peers until its queue is empty. Work peers get a unit of job from its Hub Peer and Hub Peers can get jobs at most N unit of jobs where N is the number of its work peers.

2.3 Dynamic Peer Network Structure Transformation

In P2P load distribution system, work peers' jobs are assigned from their Hub Peers and job assignment delivered to a Hub Peer is transferred through Hub Peer connection structure. Therefore the time to deliver job assignment depends on Hub Peer capacity and network connection status. Furthermore, Hub Peers who is in longer distance from the job originator need longer time to have the assigned job.

To solve this, we need new dynamic Hub Peer connection structure management scheme. Our dynamic connection structure management changes peers' connection structure based on work peers' individual capacities so that job assignment to work peers doesn't require many hops to pass. In this paper, we propose an algorithm for dynamic peer network structure transformation that can reduce the number of required Hub Peers to pass job assignment to work peers.

Our algorithm first extracts each peer's average job processing time periodically, selects one peer who has minimum average processing time, and changes the peer's peer group. The criteria relocating peer group is peers' average processing time: faster processing peer moves to faster processing peer group and slower processing peer moves to slower processing peer group. In our propose approach, faster processing peer moves closer to the center node in P2P network structure. If the peer group cannot hold new peer's joining, then the peer is sent to the next available peer group whose average processing time is close to the peer.

3 Evaluation

3.1 Simulation Environment

For simulation, we use *PeerSim* [5], which is open source Java-based P2P simulator. *PeerSim* provides both cycle-driven and discrete-event simulation, and supports for structured and unstructured P2P models. For simulation we created hierarchical peer structure and then run tasks on dynamic peer structure changes. In addition, we measured performance change when the number of peers varies. Table 1 shows parameter values for performance evaluation.

Table 1. Simulation Parameters for Dynamic Peer Network Structure Transformation

Parameters	Definition	Default Value
M	The number of peers	1,200
N	The number of work peers	200~1,000
L	The number of jobs	100,000
Cycle	Simulation cycle time	50,000
G	Network graph model	Default overlay network graph
G	Maximum broadband messages from a single peer	100
α	Peer error/leaving rate	0~20%
m	Maximum work peers for one Hub Peer	20~100
n	The number of work peers to relocate at one single cycle	20

3.2 Simulation Result

Fig. 2 illustrates network topology change from our Hub Peer selection algorithm and dynamic transformation algorithm. The network in Fig. 2 is composed 100 work peers, and we assume each Hub Peer can hold maximum 10 work peers. Left side in Fig. 2 shows initial network topology, and the result from network structure

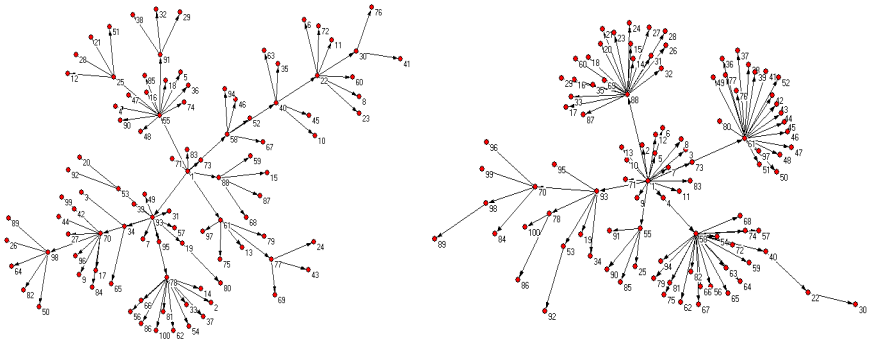


Fig. 2. Dynamic Peer Network Structure Transformation. The initial peer network topology (left), and the network topology after dynamic peer network structure transformation algorithm is applied (right).

transformation is shown in right side of Fig. 2. As you can see, all peer nodes are relocated such that the distance to center node is condensed. We used *Pajek* [6] to draw Fig. 3 from the network topology built by *PeerSim*.

By running Peer Hub selection algorithm in *PeerSim*, we created a tree structured network with various numbers of peers and applied our load distribution algorithm to assign jobs to work peers. Then we measured average processing time for each work peer. To evaluate our dynamic peer network transformation approach, we compare work peer processing time for two cases: static network topology where no peer changes its peer group and dynamic network topology where peers can change its peer group based on their average processing time.

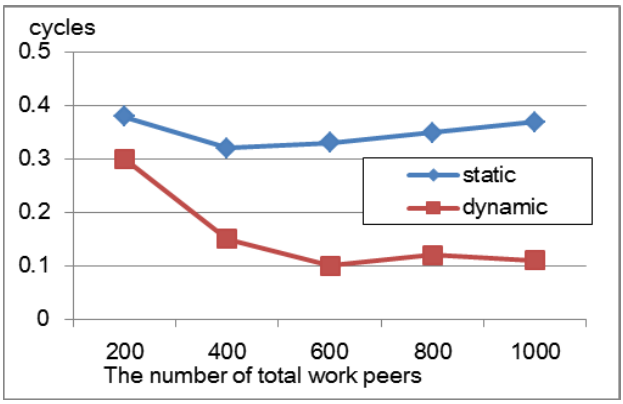


Fig. 3. Average job processing time for the number of work peers

Fig. 3 illustrates average processing time based on the number of work peers. As you can see, applying dynamic peer network structure transformation benefits in average processing time to using static peer structure. In addition, the more peers join the network, the more performance gain exists in dynamic peer network structure. Our approach also shows no distinct performance change even when the number of peers varies.

Fig. 4 illustrates average processing time based on the peer error rate. We measured performance on 1000 peers. Our dynamic peer network structure transformation approach shows better performance than static no network transformation. Error rate does not make noticeable performance changes in both approaches. Therefore, we found P2P based distributed system could work well on error-prone computing environment.

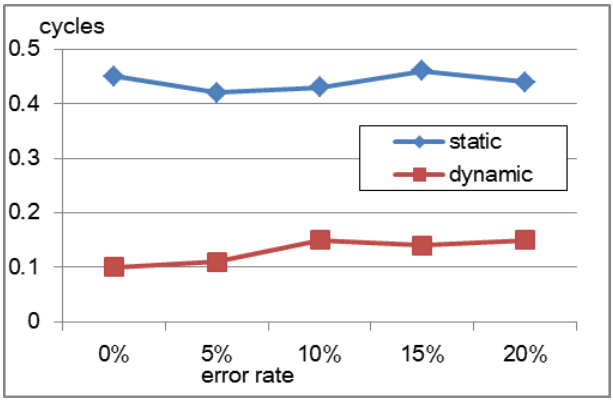


Fig. 4. Average job processing time for peer error rate

4 Conclusion

In this paper, we proposed a P2P based distribute computing model. To this end, we provide Peer Hub selection algorithm, Job distribution algorithm, and peer network structure transformation algorithm. Our approach proves reasonably good performance in term of job processing by dynamically changing peer network structure such that job is assigned to faster nodes at high priority. In addition, we found out that P2P based distributed system works well in error-prone condition. There may be some processing delay for collecting peer information and relocating peers. However, we notice such delay is negligible as there are performance benefits using dynamic network structure transformation.

Acknowledgments. This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA(National IT Industry Promotion Agency).

References

1. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum (June 22, 2002)
2. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications 15(3), 200–222 (2001)
3. Fox, G.: Peer-to-peer networks. Computing in Science & Engineering 3, 75–77 (2001)
4. Vouk, M.A.: Cloud computing — Issues, research and implementations. In: Information Technology Interfaces, ITI 2008, pp. 31–40 (2008)
5. Jesi, G.P.: PeerSim HOWTO: Build a New Protocol for the PeerSim 1.0 Simulator, `http://peersim.sourceforge.net`
6. Batagelj, V., Mrvar, A.: Pajek: Program Package for Large Network Analysis. University of Ljubljana, Slovenia (1997), `http://vlado.fmf.uni-lj.si/pub/networks/pajek`

Distributed 2D Contents Stylization for Low-End Devices

Mingyu Lim[1] and Yunjin Lee[2,*]

[1] Department of Internet & Multimedia, Konkuk University,
1 Hwayang-dong, Gwangjin-gu, Seoul, 143-701, Korea
mlim@konkuk.ac.kr
[2] Division of Digital Media, Ajou University,
San 5 Woncheon-dong, Yeongtong-gu, Suwon, 443-749, Korea
yunjin@ajou.ac.kr

Abstract. As a variety of computing devices have been developed and the Internet helps them to provide various content services in ubiquitous computing environments, users want higher quality of such services. Existing approaches focus on 3D content rendering by a remote server in order to solve the limitation of low-end devices. In this paper, we propose a distributed rendering mechanism for 2D content using multiple servers. Since large 2D image stylization also requires high computation overhead to render an image, a low-end client partition it into several image pieces. Each piece is sent to a different server, which then performs rendering. A client merges the rendered pieces to one output image again. The proposed method enables large images to be rendered by collaboration of multiple servers with reasonable processing and communication cost.

Keywords: Content stylization, Distributed rendering, Multiple servers, Low-end clients.

1 Introduction

With the advance of computing devices and Internet technologies, we can easily create and get various multimedia contents. Due to the proliferation of mobile devices which embed cameras, users have even more chance to manipulate 2D contents such as pictures, images, and video clips with a variety of applications. As recent high-end devices provides sufficient computing and rendering resources, users want to process, transform, and visualize higher quality of 2D contents.

However, normal desktop PCs and mobile devices still lack enough rendering power due to hardware limitation. To address this issue, various researches have been done using remote rendering especially for 3D contents [1]-[7]. The remote rendering method enables low-end devices to display only result images which are rendered by a remote server, but it has an intrinsic problem. Although single server has a role of surrogate renderer of a client, it can get overloaded in terms of rendering time if the server performs rendering of large contents. As the contents size increases, it takes

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_59,

longer time to process the contents. To overcome the limitation of single server, some existing approaches use parallel processing on a multi-core graphics card [8]. However, this approach requires expensive hardware and sufficient knowledge of parallel programming [9].

In this paper, we aim to distribute the rendering overhead of single server to multiple ones in order to reduce overall delay of image stylization. To this end, a client is in charge of dividing an input image into several small pieces, if the image size is greater than threshold value. Since the stylization of a pixel is affected by its surrounding pixels in our stylization method, an input image splits such that boundary pixels of pieces are overlapped. Otherwise, the stylization result at boundary pixels of two image pieces cannot be merged seamlessly. Each image piece is then sent to a different server so that multiple servers can handle only the separated image piece instead of a whole input image. The multiple servers conduct stylization of the received input image piece in parallel and independently. They do not interact with each other, and just send their result of image piece back to a requesting client. While receiving result image pieces from servers, a client assembles them into one stylized image, which is then displayed to a user. The proposed distributed stylization approach addresses the bottleneck problem of single server rendering with the modest change of client side, and makes it possible to stylize even a large input image within reasonable delay.

2 2D Image Stylization

To stylize 2D contents in our system, we use a GPU-based line drawing method proposed by Lee et al. [10]. Although this line drawing method renders a 3D mesh as a line drawing, it can be applied to any scene representation such as point set, implicit surface, or image-based representation because the method extracts lines from a shaded rendering of a scene. Therefore, it also works quite well using a 2D image as input. In addition, the method can capture the tone variations in broad region by combining toon shading with lines.

Viewing the tone image as a height field, highlight lines and dark lines correspond to ridges and valleys, respectively. To extract lines along thin areas, we apply a ridge detection method using polynomial fitting. At each pixel, we fit a degree-2 polynomial to the tone values near the pixel. In practice, we use 9 sample points arranged in a 3×3 grid around the pixel location, with spacing set to half the desired line width. We can determine if a pixel is on a ridge or valley using extracted geometric properties of a fitted polynomial.

However, since we use a degree-2 polynomial, we cannot distinguish two cases: a pixel near a ridge or valley (Fig. 1(a)) and a pixel on an edge (Fig. 1(b)). To distinguish such cases, we use an iterative search method, which moves the point sampling toward the detected ridge or valley line, fits a polynomial with new samples, and measures curvature and the distance to the new ridge or ridge line. In our implementation, the iterations are repeated at most five times. In the case (b), the new computed curvature falls below a threshold or the new location moves outside the fitted region. Details of the approach are described in [10].

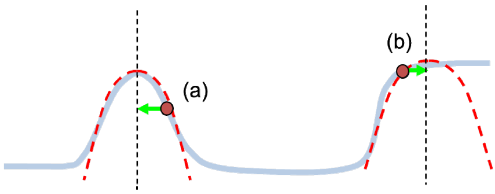


Fig. 1. Ridge searching

The line extraction process consists of two passes: the first to generate a tone image, which is a shaded rendering of a scene, and the second to detect ridges and valleys in the tone image. In the case of a 2D image as an input, the input image itself can be used as a tone image and only a blurring step is performed to reduce noise in the image or sampling artifacts in the second pass. After the line extraction process, we can augment lines with toon shading to stylize tone variations in broad regions. The whole process is performed on GPU using a fragment shader, which uses pixels only in a local region around each pixel in all passes.

3 Distributed Content Stylization

For our distributed contents stylization, we separate the required processes between a client and a server. A client is responsible for dividing an input image, merging result sub-images into one output image, and visualizing an output image. A server takes a role of main processes for stylizing an image piece received from a client.

At the client side for the distributed rendering, we need to consider how to partition an input image. In this step, we need to make the boundary regions between sub-images shared by each other in order to provide the same inputs to a pixel shader at pixels around the boundaries as the inputs in the local contents stylization.

For each pixel, a fragment shader uses 9 pixels sampled around the pixel location to fit a polynomial in the second pass of the stylization. The sampling center can move inside the initial sampling region in the iterative searching process and we sample 9 pixels around a new sampling center for each iteration. As shown in Fig. 2, if the spacing between samples is w and the initial sampling center is at (x, y), the maximum horizontal or vertical distance from (x, y) to the pixels used by a fragment shader in the second pass is 2w. As pixel values are blurred by a Gaussian kernel of size k in the first pass of the stylization, we need a $(2(2w+k)+1)\times(2(2w+k)+1)$ region centered at each pixel from an input image to determine if a pixel is rendered as a line or not.

In the image division step, we add a region of width $2w+k$ or height $2w+k$ along the boundary of each sub-image. Each sub-image is then sent to a different server and processed independently. The result sub-images from different servers are sent back to a client, and are combined into a final line drawing result of the input image.

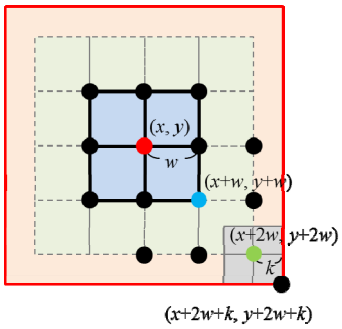


Fig. 2. A Local Region Required for the Processing

The proposed distributed stylization system is developed by separating visualization process and rendering process between a client and servers, and the communication between participating nodes are realized by our communication middleware (CM) [11]. In this section, we introduce the supporting functionalities of CM for rapid development of the distributed stylization system.

The CM aims to provide an easy and efficient way of developing distributed applications. It supports various functionalities with options for different requirements of developers. The middleware deals with multi-user issues which have to be implemented by developers if it has only fundamental networking supports. Our system has a role of bridge between an application and underlying network infrastructure. Among basics to do this is to deliver messages and contents between these two entities, by which communicating nodes can interact with each another. With APIs provided by the CM, application developers can create, send, receive, and process an event. In addition to dealing with events, it supports other operations which detect a specialized event and conduct a dedicated service according to the event type.

4 Conclusion

In this paper, we proposed a distributed rendering scheme. Using multiple servers which render and stylize only parts of a whole image, the proposed system overcomes the limitation of single rendering server, and can process high quality image with marginal processing cost. As the proposed approach uses the collaboration of distributed servers, it can be applied to cloud computing environment where multiple nodes participate in distributed stylization as one of content services.

Currently, our research is still in progress, and we are planning to conduct the quantitative performance evaluation according to different image size and number of servers in order to verify the performance. We also have a plan to extend our approach to adaptive distributed rendering technique. In this approach, the number of participating servers is dynamically chosen according to the size of input image in order to make the system more scalable than current fixed environment.

Acknowledgements. This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA(National IT Industry Promotion Agency.

References

1. Shi, S., Nahrstedt, K., Campbell, R.: A Real-Time Remote Rendering System for Interactive Mobile Graphics. ACM Transactions on Multimedia Computing, Communications, and Applications 8(3s), Article 46, 46:1–46:20 (2012)
2. Doellner, J., Hagedorn, B., Klimke, J.: Server-Based Rendering of Large 3D Scenes for Mobile Devices using G-buffer Cube Maps. In: 17th International Conference on 3D Web Technology, pp. 97–100. ACM, New York (2012)
3. Diepstraten, J., Gorke, M., Ertl, T.: Remote Line Rendering for Mobile Devices. In: Computer Graphics International, pp. 454–461. IEEE Computer Society, Washington (2004)
4. Lamberti, F., Sanna, A.: A Streaming-Based Solution for Remote Visualization of 3D Graphics on Mobile Devices. IEEE Transactions on Visualization and Computer Graphics 13(2), 247–260 (2007)
5. Paravati, G., Sanna, A., Lamberti, F., Ciminiera, L.: An Open and Scalable Architecture for Delivering 3D Shared Visualization Services to Heterogeneous Devices. Concurrency and Computation: Practice & Experience 23(11), 1179–1195 (2011)
6. Gobbetti, E., Kasik, D., Yoon, S.: Technical Strategies for Massive Model Visualization. In: The 2008 ACM Symposium on Solid and Physical Modeling, pp. 405–415. ACM, New York (2008)
7. Chang, C., Ger, S.: Enhancing 3D Graphics on Mobile Devices by Image-Based Rendering. In: The Third IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, pp. 1105–1111. Springer, London (2002)
8. Yoo, W., Shi, S., Jeon, W., Nahrstedt, K., Campbell, R.: Real-Time Parallel Remote Rendering for Mobile Devices Using Graphics Processing Units. In: IEEE International Conference on Multimedia and Expo, pp. 902–907. IEEE Press, New York (2010)
9. Nickolls, J., Buck, I., Garland, M., SSkadron, K.: Scalable Parallel Programming with CUDA. Magazine Queue – GPU Computing 6(2), 40–53 (2008)
10. Lee, Y., Markosian, L., Lee, S., Hughes, J.: Line Drawings via Abstracted Shading. ACM Transactions on Graphics 26(3) Article 18 (2007)
11. Lim, M., Kevelham, B., Nijdam, N., Magnenat-Thalman, N.: Rapid Development of Distributed Applications Using High-Level Communication Support. Journal of Network and Computer Applications 34(1), 172–182 (2011)

Authority Delegation for Safe Social Media Services in Mobile NFC Environment

Jinsung Choi[2], Okkyung Choi[1,*], Yun Cui[1], Myoungjin Kim[1], Hanku Lee[1], Kangseok Kim[2], and Hongjin Yeh[2]

[1] Center for Social Media Cloud Computing, Konkuk University, Seoul, Korea
{okchoi20,ilycy,tough105,hlee}@konkuk.ac.kr
[2] Department of Knowledge Information Security,
Graduate School of Ajou University, Suwon, Korea
popsae@naver.com, {kangskim,hjyeh}@ajou.ac.kr

Abstract. With the rapid development of NFC(Near Field Communications) technologies, NFC-enabled mobile devices are replacing the existing RFID such as mobile payment service, access control system of door locks, and ticketing service. Your service access authority through authenticating the mobile device can be delegated to any person temporarily. But when a person wants to share one's authority to others, it would be considered prevention for abuse of authority. For example, when parents give the payment right to their children, it can be used indiscriminately. And it can be abused when the authority is transferred to a third party without checking authentication code. So transferred tickets or copied access authority of door locks can occur. In this paper, for safe authority delegation, we will contain the authorized user's identity and check authorization in mobile device whether it contains suitable delegation information or not.

Keywords: Near Field Communication, Authority, Delegation, Smartphone, User Identification.

1 Introduction

With the rapid development of mobile devices such as iPhone and Android devices, the number of active smartphone users around the world has topped 1 billion according to a global research and consulting firm. And smartphone use is exponentially increasing and the firm estimates another billion will become smart phone users by 2015 [1].

IDG(International Data Group) divides NFC services into mobile payment service and other services, expecting the number of users using smartphones with NFC-enabled mobile payments to continually increase.

As shown above, NFC is used increasingly in a variety of fields but the services using mobile NFC have not yet provided services which allow the authority to use

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_60,

service to be transferred or lent to others. There is a fear that such authority may be misused or abused when parents hand payment authority over to their children or the access authority is lent to others. Therefore, it is necessary to check if the user getting access to the service has an appropriate authority so that the authority may be used in a proper manner.

This study aims to conduct the analysis of methods and issues as to authority delegation on the web which was handled in the previous paper, as well as to propose a detailed method for safe authority delegation in mobile NFC environment. Regarding the way to propose, this is intended to prevent the authority from being misused and misused by building a safe and powerful security policy so that the authority of a personal profile should be possibly delegated only for the guaranteed medium.

2 Related Work

Mambo et al.[2] is a cryptographic protocol which can create a delegated authority with the identical effect to the original authority before verifying it. But, when the delegated authority is copied or leaked to the third party, it may be misused and abused because the identification verification procedure for the person using the authority is not present. Horster et al.[3] is a protocol for generating a security key which creates a delegated key pair of the Delegate and Authority. As the Authority is verified with the used of ID and delegated key pair, the Delegate may use the Authority's ID or the Authority may pose as the Delegate with the delegated key pair. This may be referred to as misuse/abuse of authority that occurs due to unclear identification verification.

To overcome the shortcomings of the techniques for such authority delegation, Kim et al. [4] allows delegation information to include the delegation start and end dates, and the Authority and Delegate, making it unclear as to ways and procedure to check the identity of the authority user.

If identity verification is not possible when the authority is delegated and used in mobile NFC environment, it may be misused and abused by being copied recklessly to the third party. Therefore, to prevent this from occurring, it is inevitable to check if it is a mobile device which can use the delegated authority.

Park et al. [5] uses the identity verification which makes use of the certification issued by the certification authority when delegating the authority in online service environment. Both Authority and Delegate should have the certification issued for delegated registration on the server, however in mobile NFC environment this can be replaced by the mobile device information such as a device identifier without issuing the certification, and the delegated information can be even handed over from the Delegate's device to the Delegate's device through NFC (Near Field Communication).

3 Proposed Method

3.1 Overall System Flow

This study shows when the user with a delegated authority gets access to the system to use the service, the system performs the user identity verification to check if the authority owned by the user getting access to the system is appropriate. Fig.1 shows the process of an authority delegation in which the user identity verification is possible in mobile NFC environment.

Fig. 1. Overall Flow Chart

First of all, the Delegate who wishes to get the delegated authority requests a right for the service to the Authority. In the 2nd phase, the Authority and relay server carry out the key generation process for coding the authority token to be handed over to the Delegate. In the 3rd phase, the relay server stores in the database the authority delegation data received from the Authority, whereas in the 4th phase the Authority hands the coded authority token over to the Delegate. In the 5th phase, the Delegate gets NFC to contact the reader to access the service and hands over the authority token and his/her identity information. In the 6th phase, the reader checks that an appropriate authority for accessing the service is present through the comparison with the delegated data stored in the relay server database by decoding the token.

3.2 Authority Delegation Policy

Only if the Authority delegated from the Authority to the Delegate contains identifiable information, there is a fear that the Delegate may misuse the delegated authority. Therefore, instead of delegating the entire authority, partial delegation should be made by putting restrictions on authority. Table 1. is the proposed authority delegation policy.

Table 1. Authority Delegation Policy

Level	Authorization	Contents
Level1	Full Delegation	No Limitation
Level2	Periodic Partial Delegation	Limitation of period
Level3	Limited attempts Partial Delegation	Limitation of number of times
Level4	Full Partial Delegation	Can be used only once
Level5	No Permission	Service Inaccessible

The authority is rated in 5 steps. 1st grade authority involves getting all authority delegated from the Delegate, putting no limit on the period of use and number of times used. 2nd grade authority involves partial authority per period, putting restrictions in the period during which the authority can be used but putting no limits on number of times used during the period. 3rd grade authority represents partial authority per number of times used, putting limits on number of times for use of authority. 4th grade authority allows the delegated authority to be used only once. Lastly, 5th grade authority allows the use of authority to be completely restricted, including either of the following: possessing no authority as the authority was not delegated, or going beyond the restrictions where the authority owned can be used.

4 Conclusion

The recent increase in smartphone supply drives more users to utilize mobile services and it magnifies the importance of mobile security solutions. But it is not easy to defend cyber-attacks against services because the attacks get more diverse every year [6]. The method proposed in this paper enabled the user identification verification by marking restrictions on delegated authority and individual identifiers of the Authority and the Delegate before the Authority delegates authority, so as to prevent the Delegate from forging or misusing/abusing the authority at the time of delegating the authority. Unique identifiers owned by each device were used as an individual identifier, and the unique identifiers were combined to create new identifiers so that the unique identifiers should not be exposed as they stood due to a sudden possible attack. The delegation information is coded for the Delegate not to check its contents, and though the delegated authority is leaked by a malicious attack, delegation information is unable to be identified and the identity is not verified which therefore can contribute to preventing authority from being misused/abused.

Acknowledgments. This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the "Employment Contract based Master's Degree Program for Information Security" supervised by the KISA(Korea Internet Security Agency). This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA(National IT Industry Promotion Agency).

References

1. CBSNEWS, http://www.cbsnews.com/
2. Mambo, M., Usuda, K., Okamoto, E.: Proxy Signatures: Delegation of the Power to Sign Messages. IEICE Trans. on Fundamentals E79-A(9), 1338–1353 (1996)
3. Horster, P., Michels, M., Petersen, H.: Hidden signature schemes based on the discrete logarithm problem and related concepts. In: Proc. of Communications and Multimedia Security 1995. Chapman&Hall (1995)
4. Seong-Yeol, K., Dong-Hyun, K.: A Protocol to Delegate Signing Right for Multi-level Proxy Signature. J. Korean Institute of Electronic Communication Sciences 3 (2009)
5. Se-Joon, P., Yong-Jun, L., Hae-Suk, O.: Efficient Proxy Signature Technology using Proxy-Register. J. Korea Information and Communications Society 29 (2003)
6. Yunlim, K., Okkyung, C., Kangseok, K., Taeshik, S., Manpyo, H., Hongjin, Y., Jai-Hoon, K.: Two-factor authentication system based on extended OTP mechanism. J. Computer Mathematics 89, 1–15 (2013)

Introspection-Based Periodicity Awareness Model for Intermittently Connected Mobile Networks

Okan Turkes, Hans Scholten, and Paul Havinga

Dept. of Computer Engineering, Pervasive Systems,
University of Twente, PO-Box 217, 7500 AE,
Enschede, The Netherlands
{o.turkes,hans.scholten,p.j.m.havinga}@utwente.nl
http://ps.ewi.utwente.nl

Abstract. Recently, context awareness in Intermittently Connected Mobile Networks (ICMNs) has gained popularity in order to discover social similarities among mobile entities. Nevertheless, most of the contextual methods depend on network knowledge obtained with unrealistic scenarios. Mobile entities should have a self-knowledge determination in order to estimate their activity routines in a group of communities. This paper presents a periodicity awareness model which relies on introspective spatiotemporal observations. In this model, hourly, daily, and weekly locations of mobile entities are being tracked to predict future trajectories and periodicities within a targeted time period. Realistic simulations are utilized to analyze the predictions in weekly observation sets. The results show that a reasonable accuracy with an increasing level of determination can be obtained which does not require global network knowledge. In this regard, the presented model can give insights for any type of ICMN objectives.

Keywords: Intermittently-connected mobile networks, social networks, context-awareness, periodicity awareness model, spatiotemporal correlations.

1 Introduction

Opportunistic way of communications has become feasible with the idea of forming circumstantial networks such as Intermittently Connected Mobile Networks (ICMNs), with ubiquitous mobile- and smart-phone carriers [1]. ICMNs provide flexibility and scalability for information sharing, especially in case of overloaded or inoperable wired/wireless infrastructures during emergencies. Unlike conventional wireless networking approaches, ICMNs focus on exploiting the advantages and obviating the disadvantages of ever-changing mobility dynamics for data routing and dissemination. In order to form collaborative/cooperative node clusters for communicating with high efficiency in communities or connecting dispersed vicinities with suitable intermediary nodes, research motivations for ICMNs converge on context-awareness (CA) to eliminate forecast uncertainty due to unpredictable nature of mobile entities [2].

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_61,

For CA, ICMNs can exploit either self-knowledge (K_S), or vicinity-knowledge (K_V), or both to a certain extent. Most of the routing strategies rely on a full K_V, in which global network properties are known beforehand or provided routinely. Approaches, which exploit relative significance values among node locations and measure several graph metrics such as centrality and closeness, initially obtain a partial or full K_V as well. Such unreasonable scenarios may substantially cause an unexpected level of dissemination in reality [3]. With more realistic scenarios, methods such as social labeling, ranking, and temporal-based comparisons mostly rely on an encounter-based K_V. However, determining socially (in)coherent nodes among contacts may not necessarily provide a high efficiency and scalability in information sharing, since spatiotemporal node behaviors and relationships are especially evaluated for only small-scale vicinities. There are self-aware methods which focus more on K_S to determine social similarities over large-scale ICMNs [4-6]. However, they do not observe mobility routines for longer period of times. To the best of our knowledge, spatiotemporal periodicities in ICMNs have still not been investigated thoroughly.

This study presents a periodicity awareness model for ICMNs which relies on introspective reasoning of spatiotemporal trajectories. Only utilizing K_S, locations of network entities are being tracked for long period of time. Hourly, daily, and weekly sets are formed in order to record and determine regions of all of the mobile nodes at specific time intervals. Then, these spatiotemporal regions are analyzed to build a prediction set for future locations and trajectories within a targeted time. On the contrary, K_V is completely ignored, such that the level of network dependency is attained at its lowest. With an efficient and realistic introspection, simulation results indicate that a reasonable awareness can be achieved for any kind of objective in ICMNs.

The rest of the paper is organized as follows: Section 2 explains our periodicity awareness model. Section 3 demonstrates our experimental setup. Section 4 gives the performance analysis. Section 5 gives the conclusion and outlines the future work.

2 Introspection-Based Periodicity Awareness Model

The proposed periodicity awareness model relies on location and time context. We use K_S of mobile spatiotemporal periodicity data for two basic reasons: 1) Mobile nodes must have selective message switching mechanisms to decide on when to relay the information to the other network entities. 2) Message sharing methods must investigate mobile message carriers' attitude to project future trajectories or contacts.

We define two dynamic buffers, $L = \{l_1, l_2,\ldots, l_k\}$ and $T = \{t_1, t_2, \ldots, t_k\}$, to keep track of the locations and the corresponding times, respectively. The number of measurements, denoted by k, can be adjusted according to the application necessities. We define an ordered pair, $s_i=(l_i, t_i) \mid 1 \leq i \leq k, l_i \in L, t_i \in T$ to form a general spatiotemporal set $S = \{s_1, s_2,\ldots, s_k\}$. A weekly record set, $W = \{S_1,\ldots, S_7\}$, contains all recordings of a specific day (S_d) of the previous week. Every week, recordings

in the sets are updated. Analogically, we define an identical instance of W, as $\overline{W}$, to investigate the generality of the weekly recordings. According to that, average spatiotemporal pair set, that is $\bar{s}_i=(\bar{l}_i, t_i)$, is recorded in $\overline{W}$ and is also updated every week as given in Equation 1, where c demonstrates the number of observation weeks.

$$\bar{s}_i = \begin{cases} \dfrac{c \times \bar{s}_i + s_i}{c+1}, if\ W \neq \emptyset \\ s_i \qquad\quad , if\ W = \emptyset \end{cases} \tag{1}$$

Spatiotemporal data are recorded to project trajectories and, therefore, to estimate total distances (x) and displacements (ΔR) for a targeted time t_g. Starting from the current time t_c, until $t_c+ t_g$, pairs of s_{c+n} and $\bar{s}_{c+n}$ ($1 \leq n \leq g, n \in \mathbb{N}$) are obtained from the sets of the same day, S_d and $\bar{S}_d$. For t_c, we have the location from the preceding week ($l_c \in S_d$) and the average location for the weeks observed ($\bar{l}_c \in \bar{S}_d$), where the Euclidean center of them is μ_c. Then, we find the Euclidean vector v which is from μ_c to the real current location. For each subsequent s_{c+n} and $\bar{s}_{c+n}$, we find μ_{c+n}. We have the set $M = \{\mu_c, \mu_{c+1}, \dots, \mu_{c+g} \mid c, g \in \mathbb{N}\}$ at $t_c+ t_g$. For each two consecutive elements of M, we update the ratio ϕ_n between μ_{c+n} and μ_{c+n+1}. The terminal point of the vector ϕv gives us the projected location at that time, which is θ_{c+n}. We have the set $\Theta = \{\theta_c, \dots, \theta_{c+g} \mid c, g \in \mathbb{N}\}$ for the projected locations at $t_c+ t_g$. As Equations 2 and 3 show, we then calculate the estimated x and ΔR between t_c and t_g.

$$\tilde{x} = \sum_{n=1}^{g} \sqrt{(\theta_{c+n} - \theta_{c+n-1})^2} \tag{2}$$

$$\Delta\tilde{R} = \sqrt{(\theta_{c+g} - \theta_c)^2} \tag{3}$$

We also check the distance between the estimated locations $l_{c+g} \in W$ and $\bar{l}_{c+g} \in \overline{W}$ of a mobile node. As given in Equation 4, if it is shorter or longer than a threshold distance (τ), we set the periodicity degree ($^\circ\rho$) of that node to high or low, respectively.

$$^\circ\rho = \{ Low, if \left| l_{c+g} \in W - l_{c+g} \in \overline{W} \right| < \tau, \text{or } High, if\ otherwise \tag{4}$$

This periodicity awareness model gives an insight for any type of ICMN objective by estimating future locations, trajectories, and displacements. As shown in Figure 1, when two nodes meet at t_c, they calculate their own $\tilde{x}$, $\Delta\tilde{R}$, and $^\circ\rho$ for $t_c+ t_g$. If the objective is the collaboration for an event detection, they can compare their $\Delta\tilde{R}$ whether to see they stay in the same region, or not. If the objective is data dissemination over large regions, they can compare their $\tilde{x}$ to understand which one is more dynamic to carry the message to the other regions. If the objective is creating social-coherent clusters, they can check their $^\circ\rho$ and decide to involve in that community, or not. In this model, nodes utilize K_S with their own observations and do not require K_V.

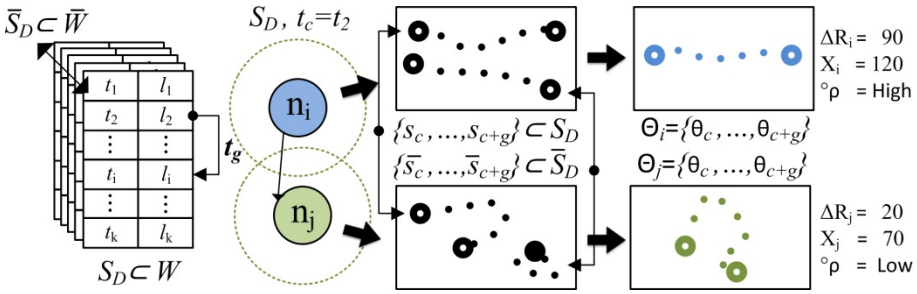


Fig. 1. The periodicity awareness model

3 Experimental Setup

The Opportunistic Networking Environment simulator [7] with the shortest path map-based movement model is used to evaluate our proposed model. In the experiments, a virtual city of approximately 5000m by 5000m is generated to analyze 500 individuals as mobile entities under 4 different groups, as shown in Table 1. In the city map, 5 different points-of-interests (POIs) types are defined with several locations. For instance, residential and commercial POIs are dispersed over the city map whereas marketplace POIs are located near to the city center. Several range of POI probabilities are assigned for each group. In the simulation, a worker can drop over places such as his/her house, office, and restaurant whereas a student can shuttle between places such as his/her house, school, and library. A housewife spends most of her time at home and usually goes for shopping. Thus, these 3 groups generally move with a purpose. Besides, a tourist is a random-stroller in the simulation.

With this scenario, we draw a parallel between the real world and the simulation setup for daily city activities. In addition, by assigning varying characteristics to the mobile entity groups, we define 5 different settings with random seed numbers in order to represent each workday of a week. Random seed numbers in the simulation create different trajectories for one-day-long activities of the mobile entities, meaning that the individuals have appraised POIs with varying schedules. Spatiotemporal data are tracked to create the sets W and $\bar{W}$, which are utilized to predict the periodicities.

Table 1. POI probabilities for different mobile entity groups in the simulation

Groups/POIs	Residential	Commercial	Educational	Recreational	Marketplace
Workers (125)	0.20-0.25	0.35-0.40	0.00-0.05	0.05-0.15	0.10-0.15
H.wives (125)	0.35-0.40	0.00-0.00	0.05-0.10	0.10-0.15	0.30-0.35
Students (125)	0.25-0.30	0.00-0.00	0.30-0.35	0.20-0.25	0.05-0.10
Tourists (125)	0.00-0.20	0.00-0.20	0.00-0.20	0.00-0.20	0.00-0.20

4 Performance Analysis

The soundness of the presented model is investigated by analyzing the spatiotemporal periodicities, estimated distances ($\tilde{x}$), and displacements ($\Delta\tilde{R}$) of mobile entities. For each entity, spatiotemporal data is recorded for 40 different experiments, which form 8 weekly observation sets by each of 5 weekday tests. The sensing interval differs between 300sec and 3600sec. Figure 2(a) shows the difference between the real and estimated displacements ($|\Delta R - \Delta\tilde{R}|$) with regard to number of weeks (c) and sensing interval, where t_g is twice of the sensing interval. It is evident that sensing frequency has a positive effect on the ΔR estimation, however, to a certain extent. Besides, an increase in c does not substantially improve the accuracy of ΔR estimation. As Figure 2(b) depicts, if mobile entities are investigated separately, the effect of periodicity on x estimation can be seen clearly. Strict periodicities generate accurate x estimations. This means that the presented periodicity model is also suitable for determining trajectories of vehicles and people with predetermined routes carrying wireless modules.

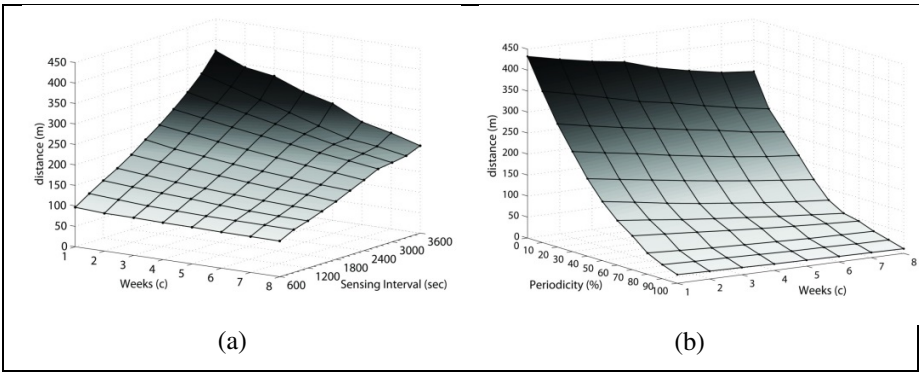


Fig. 2. The effect of sensing interval and periodicity on estimated distances

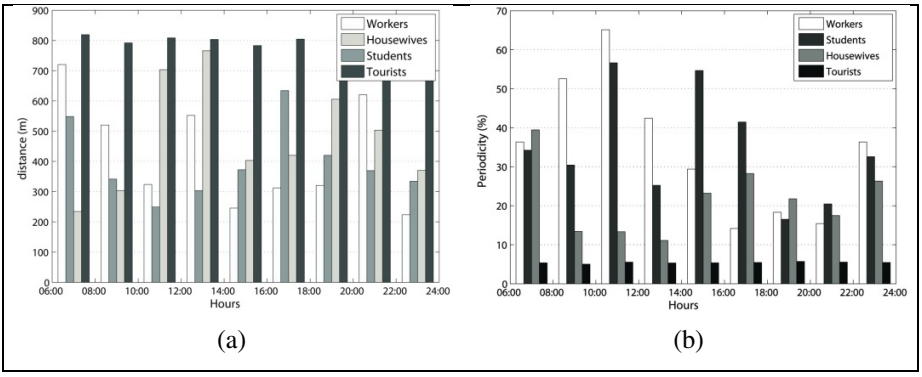


Fig. 3. Periodicities of mobile groups with respect to time and distance traveled

Figure 3(a) shows $\tilde{x}$ results for each entity groups according to different day hours when t_g=1800sec. In addition, the hourly changes in periodicities for each group are shown in Figure 3(b), by calculating the percentage of entities which stay in the same POI region within t_g among all group entities ($^{\circ}\rho = High, \tau = 500m$). By analyzing $\tilde{x}$ and $^{\circ}\rho$ results together, we can compare the social routines of the entities. Thus, our model can be suitable to fulfill any ICMN objective such as collaboration, dissemination, and clustering which are the tasks T_1, T_2, and T_3 defined in Table 2, respectively.

Table 2. Number of appropriate nodes selected at specific POIs and times for the defined tasks

Tasks/Regions (t_g=30min)	Residential	Commercial	Educational	Recreational	Marketplace
T_1: $\tilde{x} \leq$100m $\wedge$ $^{\circ}\rho$ = High	11 (04:00) 6 (20:00)	9 (11:00) 3 (16:00)	10 (11:00) 5 (16:00)	6 (09:00) 6 (17:00)	3 (09:00) 5 (17:00)
T_2: $\Delta\tilde{R} \geq$100m $\wedge$ $^{\circ}\rho$ = Low	2 (04:00) 5 (20:00)	4 (11:00) 9 (16:00)	5 (11:00) 7 (16:00)	14 (09:00) 22 (17:00)	11 (09:00) 32 (17:00)
T_3: $\tilde{x} \leq$1000m $\wedge$ $\Delta\tilde{R} \leq$500m	16 (04:00) 8 (04:00)	13 (11:00) 4 (16:00)	10 (11:00) 6 (16:00)	8 (09:00) 7 (17:00)	10 (09:00) 9 (17:00)

The number of appropriate nodes for T_1 is high for the POIs which are visited at least for once by each mobile group. For example, all of the mobile groups stay at residential areas during night. Similarly, student and worker groups provide better results for T_3 during work hours. Besides, recreational and marketplace POIs are suitable For T_2.

5 Conclusion and Future Work

In this paper, we propose a periodicity awareness model for ICMNs with introspective spatiotemporal observations. We record the locations of mobile entities with different time intervals to discover daily and weekly movement routines. We present an estimation model to predict trajectories and periodicities. The model projects the future locations in a targeted time period by utilizing weekly observation sets. Realistic simulations show that we can obtain reasonable accuracy with an increasing level of determination without global network knowledge. The proposed model gives insights for any kind of ICMN objective. As future work, the model should utilize more precision controls and knowledge for periodicity awareness. The model will be used as basis for an efficient data dissemination protocol in a network of mobile phones.

Acknowledgments. This study is supported by the SenSafety project in the Dutch Commit program.

References

1. Khabbaz, M., Assi, C., Fawaz, W.: Disruption-tolerant networking: A comprehensive survey on recent developments and persisting challenges. IEEE Communications Surveys Tutorials 14(2), 607–640 (2012)

2. Bellavista, P., Corradi, A., Fanelli, M., Foschini, L.: A survey of context data distribution for mobile ubiquitous systems. ACM Comput. Surv. 44(4), 24:1–24:45 (2012)
3. Zhu, Y., Xu, B., Shi, X., Wang, Y.: A survey of social-based routing in delay tolerant networks: Positive and negative social effects. IEEE Communications Surveys Tutorials 15(1), 387–401 (2013)
4. Hsu, W.-J., Dutta, D., Helmy, A.: Profile-cast: Behavior-aware mobile networking. In: Wireless Communications and Networking Conference, WCNC 2008, March 31-April 3, pp. 3033–3038. IEEE (2008)
5. Thakur, G.S., Helmy, A., Hsu, W.-J.: Similarity analysis and modeling in mobile societies: the missing link. In: Proceedings of the 5th ACM Workshop on Challenged networks, CHANTS 2010, pp. 13–20. ACM, New York (2010)
6. Moon, S., Helmy, A.: Understanding periodicity and regularity of nodal encounters inmobile networks: A spectral analysis. In: 2010 IEEE Global Telecommunications Conference (GLOBECOM 2010), pp. 1–5 (December 2010)
7. Keränen, A., Ott, J., Kärkkäinen, T.: The ONE Simulator for DTN Protocol Evaluation. In: SIMUTools 2009: Proceedings of the 2nd International Conference on Simulation Tools and Techniques, ICST, New York, NY, USA (2009)

Collaborative Recommendation of Mobile Apps: A Swarm Intelligence Method*

Xiao Xia, Xiaodong Wang, Xingming Zhou, and Tao Zhu

School of Computer Science, National University of Defense Technology, Changsha, P.R. China, 410073

Abstract. The explosive growth of mobile apps has given rise to the challenge of finding out interesting apps for users. Recommender systems are employed to meet this challenge. However, as the lack of user and app data, the development of recommender systems for mobile apps is still at a slow pace. Therefore, we propose a system-level collaboration approach to facilitate the development of new systems by making a better use of the data from existing systems. To this end, we model the recommendation generation as an optimization problem and propose a new set-based particle swarm optimization method to solve it. We further develop three systems to evaluate our approach and algorithm. Evaluations based on real data have verified their performances on both the effectiveness and the efficiency.

Keywords: Mobile app, recommender system, collaboration, set based PSO.

1 Introduction

As reported by recent investigations, the milestone of 700,000 apps has been hit in both the Google Play market and the Apple App Store [1]. Such an explosive growth in the population of mobile apps makes it much more difficult for users to find out apps of interests. Recommender systems thus are employed to meet this challenge. For instance, AppJoy [2] utilizes the personalized usage patterns of users to recommend apps. AppBrain [3] generates recommendations within the same category by monitoring the installation history of apps. Other studies such as AppAware [4] provide recommendations by integrating the context information of devices.

Such systems are of help to users for the app discovery. However, the development of new recommender systems for mobile apps is still at a slow pace because of the lack of user and app data. Therefore, we propose a system-level collaboration (SLC) approach to tackle this challenge. Such an approach not only integrates algorithms as traditional methods, but also utilizes the supporting data of systems. For instance, the Google Play recommends apps based on user behaviors while AppBrain based on the app category. By collaboration of the two, we are then able to recommend apps based on both the user behaviors and the app features, even though we hold neither of them. Thus our approach facilitates the development of new recommender systems.

* This work is supported by the projects of National Natural Science Foundation of China: No. 61070201, No. 61170260 and No. 61202486.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_62,

To implement the SLC, we formulate the recommendation generation as an optimization problem. To solve the problem, we address a new set-based optimization problem and design a novel set-based particle swarm optimization algorithm, i.e., the CF-SPSO. Further for evaluations, we define three mainstream recommendation objectives and pick up three popular android app recommender systems as our source systems. We then develop three new recommender systems leveraging the SLC approach and the CF-SPSO algorithm. Each of them focuses on one of the recommendation objectives. Evaluations have shown the promising effectiveness and efficiency.

2 Collaborative Recommendation

Hybrid strategies in the context of recommender systems have verified the strength of collaborative recommending approaches [5, 6]. However, traditional hybrid strategies mainly focus on combining different algorithms. Little attention has been paid to the collaboration among systems. Therefore, we propose the system-level collaboration (SLC) recommending method. The basic idea of the SLC is to collect diverse recommendations of different systems for same apps. Advanced recommendations are then generated from the collective recommendations following the system objectives.

The SLC approach overcomes a significant challenge, i.e., the lack of large scale user and app data, for developing new mobile app recommender systems. It conquers such a barrier by making use of the supporting data behind different existing systems. For instance, the Google Play market recommends apps based on user behaviors while AppBrain recommends apps within the same category of the apps which have been accessed. Through collaboration between such two systems, we are then able to generate recommendations based on both the user behaviors and app features.

3 Problem Definition

To implement the SLC approach, we formulate the recommendation generating process for each app k as an optimization problem.

$$Max \quad Objective(R(k)) \tag{1}$$

$$s.t. \quad R(k) \subseteq \bigcup_{i=1}^{N_S} R_i(k), \tag{2}$$

$$| R(k) | = N_R, \tag{3}$$

$$D_{price} \geq \lambda_p, \tag{4}$$

$$D_{category} \geq \lambda_c. \tag{5}$$

In the formulation, sets $R_i(k)$ denote the apps recommended for k by the source recommender system i. The N_S is the number of source recommender systems while the N_R defines the number of apps in the collaborative recommendations. The λ_c and λ_p are diversity control parameters, i.e., the category diversity parameter and the price

diversity parameter. They are defined to achieve better robustness of the system. To be formal, they are defined as:

$$D_{category}(R(k)) = |\{i, cate.(i) = cate.(k)\}| / |R(k)|, \tag{6}$$

$$D_{price}(R(k)) = |\{i, price(i) = 0\}| / |R(k)|. \tag{7}$$

Therefore, $D_{category}$ captures the user preference of discovering apps in same categories. D_{price} controls the proportion of free apps and the paid ones thus to balance the user preferences and the market expectations.

The *Objective* can be defined specifically for each newly developed recommend system. To conduct the evaluations, we define several mainstream objectives of recommender systems as follows, which capture both the needs of users and the expectations of app markets.

Similarity. The "similarity" is a widely adopted recommendation objective as systems usually focus on recommending apps that are similar to those have been accessed. However, there is no ground truth suggesting the similarities among apps. Therefore, we propose a new method, which measures the app similarity by measuring the similarity among their descriptions on the market. This method results from that descriptions are used not only by developers to specify their apps but also by users to understand the apps. The similarity among app descriptions thus can serve as an indicator of similarity among apps.

To obtain confident measurement of the app similarity, we introduce the *Latent Semantic Analysis* (LSA) method. It adopts the *Vector Space Model* to represent app descriptions as vectors of weighting terms, e.g., tf-idf terms. It then projects the space of terms to the space of concepts through the *Singular Value Decomposition*. Eventually, the similarity measurement utilizes the semantic concepts instead of raw terms thus it is expected to gain a more reliable understanding. To be formal, for app i and j, their similarity can be measured by the *Cosine Similarity Distance* based on their concept vectors, i.e., the $C_i = (w_{i1}, w_{i2}, ..., w_{ic})$ and the $C_j = (w_{j1}, w_{j2}, ..., w_{jc})$:

$$similarity(i, j) = \sum_k w_{ik} w_{jk} \Big/ \sqrt{\sum_k w_{ik}^2 \sum_k w_{jk}^2}, \tag{8}$$

while the dissimilarity can be denoted as $dissimilarity(i, j) = 1 - Similarity(i, j)$. Then the similarity of recommendations for an app can be defined as the average similarity between the app and its recommendations. Let $R(k)$ be the recommendations for k:

$$Similarity(R(k)) = \sum_{x \in R(k)} similarity(x, k) / |R(k)|. \tag{9}$$

Diversity. The "diversity", which indicates whether the recommended items are different from each other, has been realized as one of the key factors to satisfy users. It can help users discover apps that may not be found solely based on the users' historical experiences. As in [7], we define the diversity of recommendations as the average dissimilarity of all pairs of apps in the recommendation set. To be formal, given recommendations $R(k)$ for the app k:

$$Diversity(R(k)) = \frac{\sum_{i,j \in R(k), i \neq j} dissimilarity(i,j)}{(|R(k)|(|R(k)|-1))}. \quad (10)$$

Utility. Some systems may mainly focus on their profits when recommending apps. Therefore, the objective "utility" is defined to characterize the profit potential of recommended apps. To this end, we characterize the profit potential of an app using its price and the number of its installations. Therefore, the utility of a recommendation set is defined as the total profits of all apps in it:

$$Utility(R(k)) = \sum_{i \in R(k)} lg(price(i) * install(i) + 1) / |R(k)|, \quad (11)$$

where the operator *lg* is introduced to lower the scale of app installations.

4 Algorithm Design

To solve the recommendation generation problem introduce by the SLC approach, we present the set-based PSO algorithm in this section.

4.1 Problem Statement

As candidate solutions of the recommendation generation problem are app sets, we model it as a set-based optimization problem. However, it is distinguished from the traditional set-based problems [8]. Firstly, apps are different from traditional items. They are with neither weights as the multidimensional knapsack problem nor neighborhoods as the travelling salesman problem. Thus there is no intrinsic supporting

Table 1. Notation definitions

Notation	*Definition*
$RS_i(k)$	the app recommendations provided for app k by source RS i
$U(k)$	the union of the app recommendations for app k from all source RSs, that is $U = \bigcup_{i=1}^{N_{RS}} RS_i(k)$
$P(U(k))$	the power set of $U(k)$, that is, $X \in P(U(k)) \Leftrightarrow X \subseteq U(k)$
f	the fitness function $f : P(U(k)) \rightarrow R$, that is, $f = Objective(X)$
X_i	the position of particle i where $X_i \in P(U(k))$ and $\|X_i\| = N_R$, the j dimension is X_{ij}
S	the swarm of particles, that is, $S = \{X_i \mid 1 \leq i \leq N_S\}$
V_i	the velocity of particle i which indicates the position update, that is, $V_i: P(U(k)) \rightarrow P(U(k))$
V_{ij}	the j dimension of velocity vector V_i with $V_{ij} \in N$
t	the current iteration count of position updates
t_{max}	the predefined maximum iteration
P_i	the previous best position of particle i, that is, $f(P_i) = max\{f(X_i^c) \mid 1 \leq c \leq t\}$
G	the global best position of the entire swarm S, $f(G) = max\{f(X_i^c) \mid 1 \leq c \leq t, 1 \leq i \leq N_S\}$

information to define the relationship and conduct the substitution among them when updating candidate solutions. Secondly, as we provide a Top-N app recommendation, the sizes of our solutions are fixed, while those in traditional problems are variable.

4.2 Cylinder Filling SPSO

To tackle the above problem, we propose the CF-SPSO algorithm with notations in Table 1. The algorithm is inspired by the structure of the revolver, which holds firing chambers arranged in a circle in the cylindrical block. Following such manners, we arrange the candidate apps in a ring topology and select a fixed number of them to fill the candidate solutions. Thus it meets the special features of our problem, i.e., the app neighborhood and the fixed-size solution. Since the process looks like constructing a cylinder and fill a fixed number of bullets in it, we name the app neighborhood arrangement as cylinder construction and the candidate filtering as cylinder filling.

Cylinder Construction. This step is to assign app weights and neighborhood relationships. To find better solutions, we define the app weight based on the corresponding fitness function, i.e., *Objective(R(k))*. We then arrange the apps in a ring topology by their weights. To be formal, we define the app weight based on the objectives of the system, which are discussed respectively in Section 3. That is, for $j \in U(k)$:

$$W(j) = similarity(j,k)\ or\ diversity(\{j,k\})\ or\ utility(j) + Random. \tag{12}$$

The $Random$ is introduced in case there are apps with same weights. We then construct a cylinder with $|U(k)|$ chambers each of which is identified by a unique number in $[0, |U(k)|-1]$. We further assign apps into the cylinder according to their weight rankings in $U(k)$. To this end, we define the function $Cylinder: U(k) \to N$:

$$Cylinder(j) = Ranking(W(j)), \tag{13}$$

where $W(i) < W(j) \Leftrightarrow Cylinder(i) < Cylinder(j)$ when $i, j \in U(k)$. Moreover, we have its inverse function $Cylinder^{-1}: N \to U(k)$ to find apps by their chamber identifiers. After all, all apps are weighted and arranged into a circle topology by this step.

Cylinder Filling. This step is to fill a fixed number of bullets to the app cylinder so to pick up apps for candidate solutions. Thus the position of a particle is mapped to a cylinder filling. The position update is mapped to the cylinder refilling, which is consisted of all the *bullet refilling* operations. Accordingly, the velocity of a particle is mapped to the change of filling schemes. We define the bullet refilling operator $\oplus: U(k) \times N \to U(k)$ over the app cylinder by mapping to the replace of one app in the candidate solution:

$$\oplus(i,n) = Cylinder^{-1}(Cylinder(i) + n \bmod |U(k)|), \tag{14}$$

where $i \in U(k)$ and $n \in range(0, |U(k)|-1)$. It indicates the moving of a filled bullet from the app chamber i by distance n, which is determined by the corresponding dimension of the velocity. We also expand the operator $\oplus$ to the vector computation,

where all elements of vectors in corresponding dimensions conduct the $\oplus$ operation. Moreover, we define the app subtraction operator $\ominus(i,j): U(k)^2 \to N$ as:

$$\ominus(i,j) = Cylinder(i) - Cylinder(j), \tag{15}$$

where $i, j \in U(k)$. Such an operation is defined to compute the distance between apps in the cylinder. Based on above operators, we then redefine the position and velocity update functions based on the canonical PSO:

$$X_i^{t+1} = X_i^t \oplus V_i^t, \tag{16}$$

$$V_i^{t+1} = \lfloor \omega V_i^t + R^p \otimes (P_i^t \ominus X_{ij}^t) + R^g \otimes (G^t \ominus X_{ij}^t) \rfloor, \tag{17}$$

where R_p and R_g are vectors of random numbers uniformly distributed in $[0,\varphi_p]$ and $[0,\varphi_g]$.The operator $\otimes$ is dimension-wise, where elements in corresponding dimensions of vectors conduct the multiplication operation.

Algorithm 1. The Cylinder Filling SPSO

Require: the number of swarm N_s, the number of recommended apps in collaborative recommendation N_R, and the fitness function f
for i=0 to N_s-1 **do**
  initialize X_i with random subset of $U(k)$, $|X_i|=N_R$
  initialize V_i with random vectors, $0 \le V_{ij} < |U(k)|$
  initialize P_i with X_i
end for
initialize G with X_j whose f is the maximum of all
while $f(G) \le f(RS_i)$ or $t \le t_{max}$ or $D_{price} < \lambda_p$ or $D_{category} < \lambda_c$ **do**
  for i=0 to N_s-1 **do**
    for j=0 to N_R-1 **do**
    Update X_{ij} with $X_{ij} \oplus V_{ij}$
    Update X_{ij} with $X_{ij}+1$ when $X_{ij} \{X_{iq} | 0 \le q \le j-1\}$
    Update V_{ij} with $\lfloor \omega V_{ij} + R_{ij}^p (P_{ij} \ominus X_{ij}) + R_{ij}^g (G_j \ominus X_{ij}) \rfloor$
    end for
  Update P_i with X_i when $f(X_i) > f(P_i)$
  Update G with X_i when $f(X_i) > f(G)$
 end for
end while
return G;

Step Forward. To avoid duplicate selections, we utilize a *1-step forward* strategy to guarantee that the update of each dimension of the position would not cover the previously updated ones. To be detailed, when updating the position along its dimensions, if the app of one dimension in the new position has been selected by a previously updated dimension, we forward the bullet along the cylinder for a step to select a

new app. The process will be repeated until there are no duplicates in the new position. After all, basic steps of the CF-SPSO algorithm are described in Algorithm 1.

5 Evaluation

In this section, we conduct evaluations to measure the effectiveness and efficiency of our SLC approach and CF-SPSO algorithm. To be detailed, we pick up three well known recommender systems as our recommendation sources, i.e., systems of the Google Play, the AppBrain and the AppAware. We then leverage the SLC approach to gather app recommendations from such source systems and implement the CF-SPSO algorithm to generate new recommendations for the specific objectives. To conduct comprehensive evaluations, we have developed three recommender systems, each of which focuses on one of the three objectives.

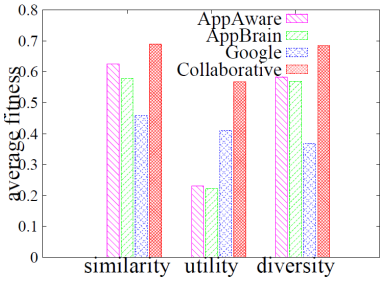


Fig. 1. Effectiveness of collaborative recommendation

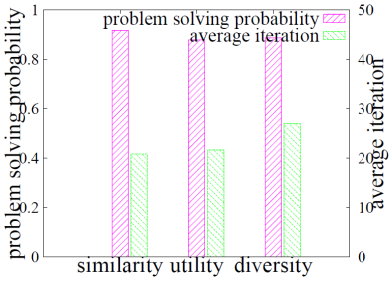


Fig. 2. Efficiency of collaborative recommendation

In the three new systems, we crawl recommended apps from all source systems for 7941 apps, the total population of whose recommendations is 108329. Using such real data, we have generated the evaluation results in Figure 1 and Figure 2, which illustrates the effectiveness and the efficiency of the SLC approach and the CF-SPSO algorithm, respectively. Figure 1 shows that for all the objectives, our approach and algorithm have generated recommendations with advanced performances, which exceed all the original recommendations from the source systems. Figure 2 tells that the CF-SPSO algorithm exhibits a promising probability of problem solving, i.e., finding out optimal solutions, and an acceptable efficiency with low average iterations.

6 Conclusion

This work indicates a new approach of developing mobile app recommender systems and identifies the potential of using swarm intelligence to recommend mobile apps.

References

1. Cnet: Google ties apple with 700,000 android apps (April 2013), http://news.cnet.com/8301-1035_3-57542502-94/google-ties-apple-with-700000-android-apps/
2. Yan, B., Chen, G.: Appjoy: personalized mobile application discovery. In: MobiSys 2011, pp. 113–126. ACM, New York (2011)
3. Appbrain: Appbrain, http://www.appbrain.com
4. Girardello, A., Michahelles, F.: Appaware: which mobile applications are hot? In: MobileHCI 2010, pp. 431–434. ACM (2010)
5. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734–749 (2005)
6. Bell, R., Bennett, J., Koren, Y., Volinsky, C.: The million dollar programming prize. IEEE Spectrum 46(5), 28–33 (2009)
7. Zhang, M., Hurley, N.: Avoiding monotony: improving the diversity of recommendation lists. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, pp. 123–130. ACM, New York (2008)
8. Langeveld, J., Engelbrecht, A.P.: A generic set-based particle swarm optimization algorithm. In: International conference on swarm intelligence, ICSI 2011 (2011)

Enhanced Implementation of Max* Operator for Turbo Decoding

Dongpei Liu, Hengzhu Liu, and Li Zhou

The Institute of Micro-electronics and Micro-processors, School of Computer, National University of Defense Technology, 410073, Changsha, China
liudongpei@nudt.edu.cn

Abstract. Max* operator is the kernel operation in MAP decoding. An intuitive approximation to the correction term of max* operator is presented. The binary-tree based architecture for multi-variable max* calculation is also suggested. The proposed max* operator provides a good trade off between hardware overhead and logic delay, and can be easily realized in parallel. Simulations on (37,21) turbo code demonstrate that the BER performance of proposed scheme is almost near the optimal Log-MAP algorithm and significantly superior to the Max-Log-MAP algorithm. The proposed enhanced implementation of max* operator has potential applications in turbo decoder.

Keywords: Max* operator, Correction function, Turbo decoding, Log-MAP algorithm, Max-Log-MAP algorithm.

1 Introduction

Turbo code provides near Shannon limit error correcting performance with acceptable decoding complexity [1]. It is widely employed in modern wireless communication systems, such as 3GPP-LTE and CDMA2000 third-generation cellular standards. The symbol-by-symbol maximum a posteriori (MAP) algorithm is the primary soft-input soft-output (SISO) decoding algorithm [2], and it can be used in the component decoders of turbo codes. The optimal Log-MAP algorithm has the merit in terms of performance. However, for high complexity of the optimal algorithm, its sub-optimal variant, the Max-Log-MAP algorithm is a compromise between performance and complexity. To improve the performance of Max-Log-MAP turbo decoders, several attempts have been done [3]-[9]. A new iterative SISO decoding algorithm based on the Viterbi algorithm was proposed in [3]. It is less complex than the Log-MAP algorithm while the performance close to it. Alternative approaches which combined the advantages of Max-Log-MAP and Log-MAP decoding algorithms were proposed in [4],[5]. In [6], [7], [8], [10], simplified correction terms, such as Linear Log-MAP, Linear-const Log-MAP approximations were presented. By contrast in [9], a constructive algorithm based on geometric programming was introduced in order to find the best piecewise linear approximation.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing,* Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_63,

In this paper, we propose an enhanced implementation of the max* operator for turbo decoding. The proposed scheme is based on an intuitive piecewise linear approximation, and can be implemented efficiently by maximum comparison in parallel. Simulations on (37,21) turbo code show that it can significantly improve the performance of the Max-Log-MAP algorithm at a cost of acceptable complexity.

2 Log-MAP Decoding Algorithm

The Log-MAP decoding algorithm for turbo codes is performed in an iterative manner. The decoder soft output, or log-likelihood ratio (LLR), corresponding to u_k is defined as [2]

$$\begin{aligned} L(u_k) = &\max^*_{(s',s)u_k=+1}\{\alpha_{k-1}(s') + \gamma_k(s',s) + \beta_k(s)\} \\ &- \max^*_{(s',s)u_k=-1}\{\alpha_{k-1}(s') + \gamma_k(s',s) + \beta_k(s)\} \end{aligned} \tag{1}$$

where $\alpha_k(s)$ and $\beta_k(s)$ are the forward and backward state metrics, respectively. After a proper initialization, the forward and backward recursion can be expressed as

$$\alpha_k(s) = \max^*_{s'}\{\alpha_{k-1}(s') + \gamma_k(s',s)\} \tag{2}$$

$$\beta_k(s) = \max^*_{s'}\{\beta_{k+1}(s') + \gamma_{k+1}(s,s')\} \tag{3}$$

where $\gamma_k(s',s)$ is the branch metric representing the logarithm of branch transition probabilities, and the max* operator (Jacobian logarithm) in the above equations (1)(2)(3) is defined as

$$\begin{aligned} \max^*(x_1,x_2) &= \log(e^{x_1} + e^{x_2}) \\ &= \max(x_1,x_2) + \log(1 + e^{-|x_1-x_2|}) \\ &= \max(x_1,x_2) + f_c(|x_1 - x_2|) \end{aligned} \tag{4}$$

The only difference between Log-MAP and Max-Log-MAP algorithms is the calculation of the max* operator. In addition to maximum comparison, a look-up table (LUT) of correction values is usually employed when Log-MAP algorithm is used in practical. If LUT is ignored, then the Log-MAP simplifies to the Max-Log-MAP algorithm at the expense of performance degradation.

3 Proposed Enhanced Implementation of Max* Operator

The max* operator greatly influences the bit error rate (BER) performance. It is calculated by maximum of two arguments plus a nonlinear term, known as correction function $f_c(x)$. Fig. 1 depicts the plot of the correction function.

An intuitive way is approximates the nonlinear correction term as a piecewise linear-constant function, in our work, we adopt the function

$$f_c(x) = \begin{cases} -0.25x + 0.625 & 0 \leq x < 2 \\ -0.0625x + 0.25 & 2 \leq x < 4 \\ 0 & x \geq 4 \end{cases} \tag{5}$$

The proposed approximation of the correction function is also plotted in Fig. 1. Note that the computation of Eq. (5) is equivalent to computing

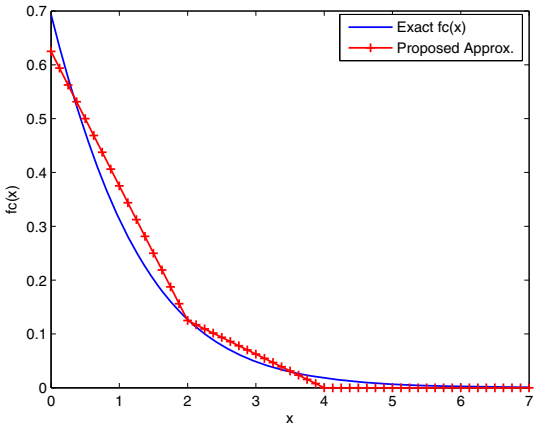


Fig. 1. The proposed approximation to the correction term

$$f_c(x) = \max(-0.25x + 0.625, -0.0625x + 0.25, 0) \tag{6}$$

Eq. (6) indicates that the proposed correction term can be efficiently implemented by parallel comparison operations. Its piecewise linear-const components can be easily realized by shift, addition operations in digital circuit which get rid of the LUT. Fig. 2 provides the VLSI architecture of the proposed max* operator. As shown in Fig. 2, the enhanced max* operator uses two shifters, five adders and three multiplexer.

Another contribution of this paper is that the multi-variable max* operator is presented as a binary-tree structure rather than traditional recursive manner proposed in [7]. Take $n = 8$ as an example, as shown in Eq. (7) and Fig. 3. The transformation in Eq. (7) indicate that multi-variable max* operator in Log-MAP algorithm can be realized in parallel, and each max* operator can be implemented independently. Therefore, it is more flexible to trade off requirements between performance and complexity.

$$\begin{aligned} &\max^*(x_1, x_2, \cdots, x_7, x_8) = \log(e^{x_1} + e^{x_2} + \cdots + e^{x_7} + e^{x_8}) \\ &= \log(e^{\log(e^{x_1}+e^{x_2})} + e^{\log(e^{x_3}+e^{x_4})} + e^{\log(e^{x_5}+e^{x_6})} + e^{\log(e^{x_7}+e^{x_8})}) \\ &= \max^*(\max^*(x_1, x_2), \max^*(x_3, x_4), \max^*(x_5, x_6), \max^*(x_7, x_8)) \end{aligned} \tag{7}$$

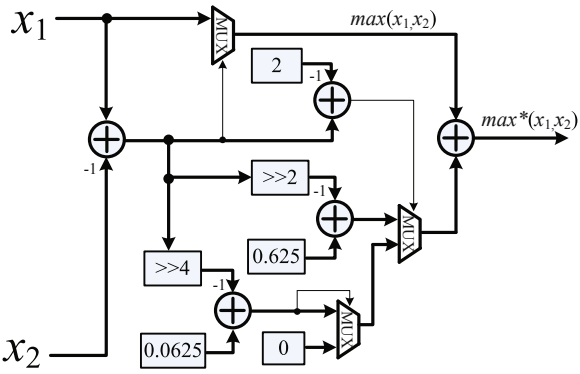


Fig. 2. VLSI architecture of the proposed max* operator

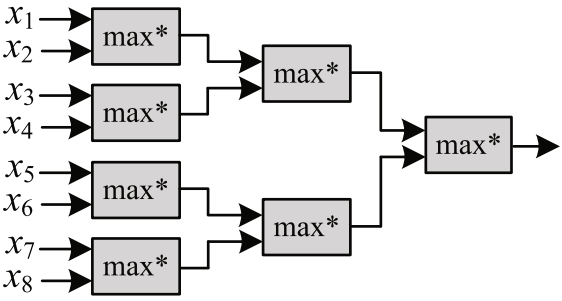


Fig. 3. The binary-tree structure for eight arguments max* computation

4 Experiment Results

In order to evaluate the hardware overhead of the max* operator, we modeled the 8-bit fixed-point max* operator and synthesized with Xilinx ISE FPGA tool chain. The proposed max* operator and some other improved max* operator are both considered. Table. 1 lists the logic area and delay of different improved max* operators for comparison. The logic utilization of the enhanced max* operator in XC5VLX110T FPGA chip is 33 slice LUTs, and the maximum combinational path delay is 6.714ns. On the one hand, the occupied resource of the proposed max* operator outperforms the Log-MAP algorithm and the method proposed in [9]. On the other hand, the combinational path has comparable delay with the method in [8] and outperforms the Log-MAP algorithm and [8], [10]. In other words, the proposed max* operator provides a good trade off between hardware cost and logic delay.

The BER performances of different algorithms are also evaluated, as shown in Fig. 4. Simulations were carried out with coding rate 1/2 turbo encoder.

Table 1. Hardware implementation of various improved max* operator

Algorithm	max* Approx.	Slice LUTs	Delay (ns)
Log-MAP	$\max^*(x_1, x_2) = \max(x_1, x_2)$ $+\log(1 + e^{-\lvert x_1 - x_2 \rvert})$	39	8.080
Max-Log-MAP	$\max^*(x_1, x_2) = \max(x_1, x_2)$	17	5.002
Linear Log-MAP Ref. [8]	$\max^*(x_1, x_2) = \max(x_1, x_2)$ $+\max(0, \ln 2 - 0.5 * \lvert x_1 - x_2 \rvert)$	30	6.715
PWL Approx. Ref. [9] (r=3)	$\max^*(x_1, x_2) = \max[x_1, x_2,$ $0.5 * (x_1 + x_2 + 1)]$	42	7.347
Linear Const Log-MAP Ref. [10]	$\max^*(x_1, x_2) = \max(x_1, x_2)$ $+f_c(\lvert x_1 - x_2 \rvert)]$	30	7.097
Proposed Algorithm	$\max^*(x_1, x_2) = \max(x_1, x_2)$ $+\max(-0.25x + 0.625,$ $-0.0625x + 0.25, 0)$	33	6.714

The generator polynomial is (37,21) in octal form, denoting the backward and feed-forward polynomials, respectively. An information sequence of 1000 bits is assumed to be transmitted over an additive white Gaussian noise (AWGN) channel. At the receiver, a pseudo-random interleaver is adopted, and the maximum iteration number is set to eight.

For the tested binary turbo codes, the turbo encoder state is sixteen for (37,21) turbo code, the $L(u_k)$ calculation is similar to the approach shown in Fig. 3,

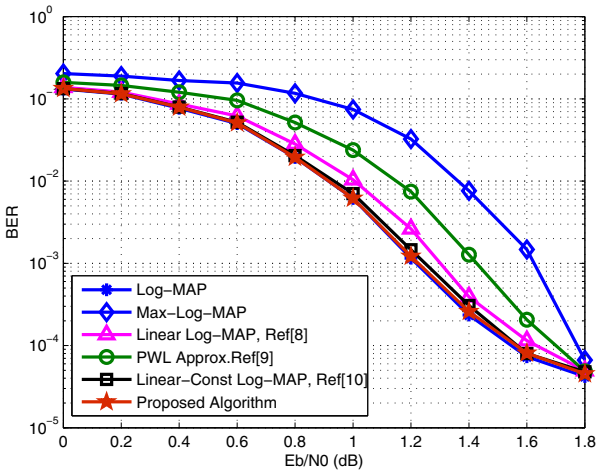


Fig. 4. BER performance of different algorithms with (37,21) turbo code

i.e., all max* operators are realized in parallel by the proposed enhanced max* operator. In Fig. 4, the BER performance of the proposed decoding scheme is very close to the Log-MAP algorithm and slightly outperforms other improved decoding algotihm [8],[9],[10]. The coding gain is about 0.30dB compared to the Max-Log-MAP algorithm at BER of 10^{-4}. Table. 1 and Fig. 4 also show, the hardware cost and logic delay of the Max-Log-MAP algorithm are both attractive, however, the performance is inferior to other improved Log-MAP algorithms.

5 Conclusion

A novel approximation to the correction term for turbo decoding is proposed, and the binary-tree structure for multi-variable max* calculation is also presented. The proposed max* operator provides a good trade off between hardware overhead and logic delay. Simulation results show that the proposed scheme has significant performance improvement compared to the Max-Log-MAP algorithm with little increased complexity. The proposed enhanced max* operation offers practical implementation advantages and has potential applications in turbo decoders.

Acknowledgements. This work is supported by the National Natural Science Foundation of China, under Grant No.60970037, and Doctor Program Foundation of Education Ministry of China, under Grant No. 20114307130003.

References

1. Berrou, C., Glavieux, A., Thitimasjshima, P.: Near Shannon Limit Error-correcting Coding and Decoding: Turbo-codes. In: Proceedings of IEEE International Conference on Communications (ICC), Switzerland, pp. 1064–1070 (1993)
2. Viterbi, A.J.: An Intuitive Justification and a Simplified Implementation of the MAP Decoder for Convolutional Codes. IEEE Journal on Select Areas in Communications 16(2), 260–264 (1998)
3. Tan, J., Stuber, G.L.: New SISO Decoding Algorithms. IEEE Transactions on Communications 51(6), 845–848 (2003)
4. Park, S.J.: Combined Max-Log-MAP and Log-MAP of Turbo Codes. Electronics Letters 40(4), 251–252 (2004)
5. Papaharalabos, S., Sweeney, P., Evans, B.G.: SISO Algorithms Based on Combined Max/Max* Operations for Turbo Decoding. Electronics Letters 41(3), 142–143 (2005)
6. Cheng, J.F., Ottosson, T.: Linearly Approximated Log-MAP Algorithms for Turbo Decoding. In: Proceeding of IEEE Vehicle Technology Conference (VTC), Spring, Japan, pp. 2252–2256 (2000)
7. Wang, H., Yang, H., Yang, D.: Improved Log-MAP Decoding Algorithm for Turbo-like Codes. IEEE Communmunications Letters 10(3), 186–188 (2006)

8. Talakoub, L., Sabeti, L., Shahrrava, B., Ahmadi, M.: An Improved Max-Log-MAP Algorithm for Turbo Decoding and Turbo Equalization. IEEE Transactions on Instrumentation and Measurement 56(3), 1058–1063 (2007)
9. Papaharalabos, S., Mathiopoulos, P.T., Masera, G., Martina, M.: On Optimal and Near-optimal Turbo Decoding Using Generalized Max* Operator. IEEE Communications Letters 13(7), 522–524 (2009)
10. Samadian, H., Nia, A.M.: Linear-constant Log-MAP, a Fast Accurate Algorithm for MAP Decoding. Journal of the Franklin Institute, 1721–1733 (2012)

A Context Description Language for Medical Information Systems

Kurt Englmeier[1], John Atkinson[2], Josiane Mothe[3], Fionn Murtagh[4], and Javier Pereira[5]

[1] Faculty of Computer Science, Schmalkalden University of Applied Science, 98574 Schmalkalden, Germany
k.englmeier@fh-sm.de
[2] Department of Computer Sciences, Universidad de Concepción, Concepción, Chile
atkinson@inf.udec.cl
[3] Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, 31062 Toulouse, France
mothe@irit.fr
[4] Department of Computer Science, Royal Holloway, University of London, Egham, England
fionn@cs.rhul.ac.uk
[5] Facultad de Ingeniería, Universidad Diego Portales, Santiago, Chile
javier.pereira@udp.cl

Abstract. Contextualized delivery of information is one of the many strengths of ubiquitous computing. It makes information actionable and helps us to better understand our situations. In the realm of healthcare, contextual information provides a terse but precise picture of the patient's health situation. The patient context can have many facets, ranging from nutrition context over health heritage context to the context of symptoms, just to name a few. Setting up the correct health condition context of a patient favors better and faster recognition of the patient's actual health situation.

Context-awareness in medical monitoring mainly concentrates on gathering numerical facts depicting special aspects of a person's health condition. In this paper we want to broaden the focus on the textual dimension in context development, by considering semantic annotation in designing context-awareness. We describe an approach for a context description language (CDL) that supports the uniform presentation of textual facts in medical reports and automatic reasoning on these facts. Term clusters in medical reports represent in a unique way symptoms and findings that set up the health context reflected in this particular report. These clusters manifest potential health condition contexts where a patient can be viewed in. A reasoning engine operates on these context presentations and selects those that match best the patient's health situation. Locating the right context supports the physician in faster getting a first picture of the probable health situation of a new patient to be examined. We present experiments with a CDL applied on reports related to respiratory problems.

Keywords: Context-awareness, context design and development, semantic annotation, domain-specific language, information mining, natural language interaction, medical reports.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_64, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

Context can be considered as a collection of facts and their inter-relationships describing the environment of a user or an event [1]. Context awareness sharpens relevance of these facts along particular aspects. It sharpens the awareness of specific restaurants in the physical context of the user, when the user is looking for one of her favorite types of places to eat. Depending on the user's preferences specific objects of the user are drawn into the context of the users while others are pushed away. In this example, it's quite likely for the user to find a restaurant of her taste. If we draw the right objects into a given context, a terse but precise picture of a particular situation is obtained. This picture helps us to better understand that situation and to make correct decisions. Consequently, it also supports disambiguation, i.e. considering an object within its correct context and excluding irrelevant contexts. Disambiguation is important in the realm of medical information where a small number of observations and/or symptoms must be brought into the correct context.

The interpretation of a health phenomenon is formed by observations and experiences in form of patients' health records. In particular, experiences include reasoning over observations and the resulting conclusions. The situation-adequate compilation of information can be essential for the appropriate handling of any health-related situation. Contextual information is used not only to interpret available information, but also to seek additionally relevant or missing information. Contextual delivery of information makes communication more efficient, more focused on the aspects of the problem at hand. It can provide information on an interesting restaurant or building the user is probably looking for [2] or it alerts persons suffering from asthma when they enter an area where others with the same health problem previously had an asthma attack [3]. In both cases, users get focused information that takes into account their personal characteristics and qualities of their immediate environment.

The two examples show that context emerges from underlying concepts and their attributes (qualities). Correctly setting up a context depends on the correct composition of these situation-specific concepts. The context "preferred restaurant nearby" or "area reportedly causing respiratory problems" is based its own theory describing a model that merges personal characteristics (preferred restaurants or chronic respiratory conditions, respectively) and with related objects at the physical location of people [4]. In the case of "asthma alert", data may come from remote monitoring of people with the same respiratory problems, i.e. monitoring the location where they used their rescue inhaler.

While the use of the rescue inhaler (location and time) can be monitored quite easily, it is a bit more complicated to represent chronic respiratory conditions. Medical reports typically provide facts representing health conditions in mainly textual form. Of course, these reports refer to a variety of symptoms that furthermore differ gradually. They represent thus variations of health condition concepts (e.g. "severe bronchial asthma" or "light bronchial asthma").

In this paper, we present a context description language (CDL) for the medical domain. With our approach we want to draw attention to semantic integration of textual facts into situation-specific contexts. We demonstrate how the CDL forms a

meta-language that supports (1) the development of context-related controlled vocabularies and (2) concept-based reasoning, which, in turn, controls the integration mechanism. Our experiment with reports on respiratory health problems exemplifies how context will be described and how a context is selected in accordance to an individual patient's health situation.

2 Natural Language Representation of Context

In the recent past, mobile computing brought us a large number of location-based services thanks to the popularity of GPS [5]. Many context engines today are built for location-based services using inference models that operate on geographical data and related mass data [6]. In automotive systems, the context engine takes the location of the car, projects the car's trajectory, and compares this path with the position or trajectory of objects in the car's way. It infers a potential collision from these data and alerts the driver correspondingly. In our case, we have to "locate" a person within the most appropriate health condition context. This helps the physician to compare the suggested contexts with the patient's health condition in order to get a solid first glance on the patient's situation and to find clues for further examination steps.

A context picture can be composed of purely numerical data or a combination of numerical and textual data. In health condition contexts, prevail textual facts interspersed with a few numerical facts. Reasoning is thus based more on methods for the semantic analysis of textual information. Our approach roots thus in Semantic Annotation that deals with analysis of unstructured content. By adding annotation tags we describe the meaning behind terms. If this description takes standardized form, machines can easier recognize facts than be solely applying text retrieval algorithms. Semantic Web knows numerous approaches for the description of content elements and relationships among them. RDF [7] is a popular standard to add structured annotations to text. It helps to bring text terms in certain semantic relationships, which supports understanding of text terms, to a certain extent, but does not solve the problem of understanding the tags. This analysis, in turn, benefits from a uniform presentation of the facts in a text. RDF does little to sufficiently address the understanding of meaning [8]. A solution to this problem provides the agreement on fixed terms to be used in text annotation. The Dublin Core Metadata Initiative[1] developed a number of concept definitions, i.e. definition of terms with attributes and properties. However, Dublin Core does not cover medical concepts. Nevertheless, a practical approach to solve the problem of understanding of meaning (in texts as well as context descriptions) emerges from the agreement on a relatively simple and precisely specified language. If this language is, in addition, machine-processable, we can develop engines that reason on context descriptions and the semantic closeness of contexts. Usually, this is the point where ontologies come into play. They define logic-based knowledge-representation formalism and are thus a powerful candidate when it comes to define context and reasoning processes to support context-awareness. OWL [9] is currently the most

[1] dublincore.org

prominent candidate among ontology languages. However, ontologies and their design languages are not only powerful but also complex. The application of ontologies, even promising, is too labor-intensive and expensive in many cases.

We therefore opt for domain-specific languages (DSL) [10] that are domain-related markup languages fostering the uniform presentation of semantics related to textual facts [11-12]. They manifest an agreement on language elements for semantic tagging and a couple of reasoning rules. Description logics expressed in DSL are not as powerful and scalable like ontologies. However, in a thematically not too complex environment, they are just as useful. Our CDL is a DSL that semantically depicts a patient's health context. The development of our CDL is guided by the assumption that humans apply a specific language in their working environment, and this becomes essentially a formal working jargon. Within this domain and community framework, natural language statements represent literal meaning [13]. Literal meaning can be interpreted correctly in the absence of any explicit and implicit context. When describing objects or processes of their working environment, users apply a language in their statements that aims at unambiguous understanding of those descriptions. The same holds for medical personnel: they want to express literal meaning in order to assure that their statements are interpreted correctly by their colleagues.

Literal meaning, in turn, is almost immediately processable by machines without language transformation that is usually required for the interpretation of (natural language) statements representing non-literal meaning that often refers to implicit knowledge of the broader statement context. In sentences representing literal meaning, there is a close correspondence between the term patterns and their inherent semantics. Furthermore, these patterns also reflect computational representations supporting taxonomic, temporal, and spatial reasoning. Text patterns have thus a clear relationship to their inherent meaning. The typical problem of ambiguity that usually comes with natural language analysis is by far less severe in statements representing literal meaning. Consequently, the logic inherent in Named Entity Recognition and some more text pattern recognition algorithms combined with a sparse semantic annotation suffice to provide enough potential to represent meaning in medical reports.

Our reasoning algorithms constituting text pattern analysis can therefore operate immediately on the users' statements. However, instead of traditional text retrieval methods [14] we propose grammar-free, machine-processable description logic [15] that assigns specific concepts to term (text) patterns within or across sentences. These concepts, in turn, form a controlled vocabulary that semantically covers the language elements for the context (model) descriptions. In order to find a proper health condition context for a particular person, a new patient for instance, we need a basic set of context descriptions (expressed in terms of the controlled vocabulary) that act as candidates for the context of this patient. From this perspective, context matching resembles information mining in unstructured data. We compare statements about an individual patient with the context descriptions in our database.

3 Representation of Context

Our CDL provides, at first, abstract concepts for both, the semantic annotation of medical reports and the abstract representation of condition contexts (Figure 1). A health condition or problem usually comprises one or more symptoms. These generic concepts represent a symptom described by (one or more) body parts (organs) and (one or more) phenomena occurring in the context of diseases (congestion, dilation, etc.). Each of these elements usually has qualities like "normal" or "calcified". In particular, phenomena have as quality also their specific location ("anterior rib arc"). The objective of text analysis is to classify all terms in the text according to these abstract concepts and to group them in accordance with their appearance in the individual reports. It is important to note, that each symptom is related to one body part and each report addresses one individual health condition.

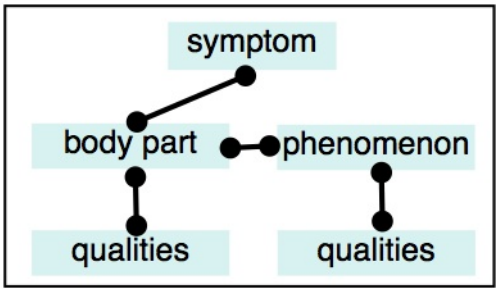


Fig. 1. Generic concepts describing health conditions. All text terms are classified along these abstract concepts.

Classification and grouping are the next features of the CDL. They are achieved by a stepwise approach (bottom-up) starting with the most simple language constructs (step 1) to more complex ones (step 5). For determining the more simple language elements we apply the well-known approach of "Parsing and Standardization" that uses stereotypical (standardized) text patterns to identify (simple and more complex) named entities.

1. Identification of Elementary Expressions

For the identification of basic elements we apply Named Entity Recognition (NER) tasks by using a set of generic patterns for Named Entities. These entities are expressions like word, number, date, or ID number. Any identified Regular Expression ("13 mm", "03-JAN-1996") here represents an instance of a basic element. Instances of these elements stand for basic concepts like word, date, number, distance, name, etc.

2. Identification of Minimal Semantics

Minimal semantics focus on generic text elements (or basic elements) and terms adjacent to these elements. The concept "age", for instance, is represented by a generic element consisting of a pattern reflecting a 3-digit Regular Expression

("1?[0-9]{1,2}") combined with terms like "age", "old", "years" or "hours". An instance of "price" may have a leading symbol ("€" or "$"); the expression "hrs." may follow an expression of digitals indicating that this element is instance of "hour".

In our approach, stop words are not eliminated, as they can be significant for the correct interpretation of entities. For identifying minimal semantics, we used n-grams that may be interlaced by stop words. To identify composed entities only the n-grams are used without their interlaced terms. These are used only for the purpose of further specification. For instance, "... valid until 30-NOV-2011" is an expression of date, represented by a 2-gram ("valid" and date) ignoring the term "until" in the first place. This term is used later when the concept "validity period" is expanded towards "end of validity period". An instance of the distance element ("21 mm") becomes a "diameter" element when followed by an expression like "in diameter". Otherwise it just represents the size of an object.

3. Transformation into Standardized Forms

Natural language is ambiguous by nature even if we consider a quasi-formal language like the one medical professional apply in their reports. Humans tend to apply different words to the same concept. We handle this synonym problem by standardization, i.e. mapping different terms representing one concept onto one representative term for this concept ("caliber" to "size" or "anterior" to "front"). It also includes the transformation of nouns into their corresponding adjective if used in order to qualify another noun like in "Aortic wall with calcification". This is transformed into "Aortic wall calcified".

Negations require a special treatment in any automatic text analysis. We treat negation as a special quality of an object. A negation just states if a quality of an object or the object itself is present or absent. In medical reports, it is important to explicitly mention the absence, for instance, of a certain phenomenon ("No pleural effusion") in order to depict a correct picture of the health condition or distinguish two otherwise similarly described conditions.

4. Identification of Key Nouns

Named entities reflect basic language elements. These expressions are input to further pattern analysis. We repeat the identification of terms frequently appearing in close proximity. This process leads to language elements that stand for concepts representing represent parts of the body ("lungs", "bronchi", etc.) or phenomena like "dilation". In the case of medical reports, nouns usually reflect these concepts.

For the development of our controlled vocabulary we extracted instances of the generic concepts mentioned above. Instances of body parts are then clustered into concepts representing larger parts of the body ("lung" comprises "right lobe", "trachea", etc.). Here, classification systems like the International Classification of Diseases (ICD) of the WHO [16] provide useful sources to set up a concept structure for body parts.

5. Conceptualization

The final step identifies key patterns as a combination of basic language elements with key nouns. This process adds qualities to concepts represented by key nouns by relating key nouns to simple language elements. Both parts usually appear within the

same phrase. A concept represented by key nouns is thus considered in a broader context including terms that reflect qualities. By using n-grams we ensure that only terms in close proximity are considered in a single conceptualization.

For example, in the phrase “Trachea and bronchi permeable, both normal caliber” we identify “Trachea” and “bronchi” as key nouns representing body parts with the qualities “permeable” and “size:normal”. As already mentioned, we apply also a normalization process that transforms, for instance, all size-related expressions into one standardized form “caliber” to “size”.

It is important to deal with pronouns and similar relationships, because they may address qualities that need to be assigned to more than one noun. In the example phrase here, the qualities mentioned address the two body parts. The keyword “both” explicitly indicates the pronoun relationship between the qualities and the two terms reflecting the body parts. It is the simple sequence of terms that determines the word pattern and thus the role of the terms (key noun - “and” - key noun - indicator of a pronoun relationship - quality-related terms) appearing in the phrase. Similar indicators include “all”, “that” or “which”.

```
<observations>
    <trachea quality="permeable">
        Trachea and bronchi permeable, normal caliber. Secretions in the trachea.
        <size>normal</size>
        <secretion/>
    </trachea>
    <bronchi quality="permeable">
        Trachea and bronchi permeable, normal caliber.
        <size>normal</size>
    </bronchi>
    <lungs>
        <consolidation quality="none">
        <granuloma quality="calified">
        <dilatation>
            Cylindrical bronchial dilatation in anterior segments of both upper lobes,
            middle lobe and posterior segments of lower lobes.
            <location>
        </dilatation>
        <surface quality="normal">
            Normal pleural surfaces.
        </surface>
        <aorta quality="calified">
            Aortic wall calcification.
        </aorta>
    </lungs>
    <heart>
        <size>normal</size>
    </heart>
    <mediastinum>
        <lymph_nodes quality="unchanged" />
    </mediastinum>
    <lump quality="benign">
        In the anterior arch of the third left rib nodular lesion shows a well defined
        contours of 13 mm diameter, nonaggressive aspect.
        <size>
            13 mm
        </size>
        <location>
            <rib quality="left,third, front" />
        </location>
    </lump>
</observations>
<conclusions>
```

Fig. 2. Medical report expressed in terms of the CDL

A pattern recognition engine takes the CDL, analyzes the medical reports and transforms them into an XML representation. The behavior of the engine is determined by the role descriptions of the concepts listed in the CDL. When applied to medical reports the CDL, the result of the analysis looks like in Figure 2.

The example shows the observation part of the report listed in the XML structure. This part is usually followed by a conclusion part (shown in Figure 3). The section shows minimal semantics ("size", "location" and "quality"), and keywords for body parts ("lungs", "aorta", "heart", etc.) and phenomena ("lump", "dilation", etc.). It also shows some results of the normalization process that standardizes "nonaggressive" to "benign" (lines 43 and 44) when appearing in proximity to the concept "lump" (lines 43 though 52) and "anterior" to "front" (lines 44 and 50).

The first phrase in line 4 implicitly contains a pronoun relationship that semantically links the qualities "size" and "permeable" to the two body parts. The representation of the two related concepts ("trachea" and "bronchi") is thus replicated.

```
54 <conclusions>
55     <bronchiectasis>
56         Bilateral bronchiectasis
57         <location>bilateral</location>
58     </bronchiectasis>
59     <atherosclerosis>
60         Aortic atherosclerosis
61         <location>aortic</location>
62     </atherosclerosis>
63     <lump>
64         Nodular lesion of the anterior arch of the left third rib.
65         <location>
66             <rib quality="left,third, front" />
67         </location>
68     </lump>
69 </conclusions>
```

Fig. 3. Conclusion section of the medical report, expressed in CDL

4 Experiments on Context Reasoning from Textual Facts

Based on a reasoning theory, we develop context engines [17] that sift through data in order to find appropriate context clues. These clues are handed over to a recommender system or, like in our case, a dialog feature presenting the clues to the user, who may then select the most appropriate context clue. We treat context here as a lexical theory more than a logical theory.

Our context engine produces a number of recommendations based on the facts collected by the physician. We apply a recommendation mechanism [18], in which recommendations are made on objects that are close to the user's context. Furthermore, the physician can gradually add more facts to enlarge proximity. We take this approach, because the first hints on a patient's situation usually lack accuracy. By viewing recommendations and gradually narrowing in the number of potential recommendations the physician's view gains accuracy.

We ran an experiment with 170 medical reports in order to see how a CDL can be used to set up a specific controlled vocabulary and to generate context instances for respiratory health problems. An average report has a length of 250 words and mentions 12 symptoms. All reports address respiratory problems. The symptoms described in the reports, however, refer to lungs and respiratory tracts but also to adjacent organs like the heart.

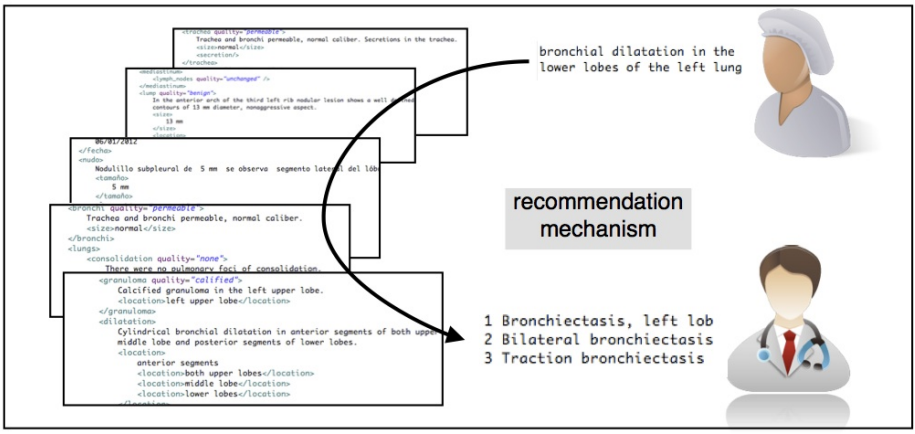


Fig. 4. The recommendation mechanism takes the patient description, locates the most appropriate contexts among all available health condition contexts, and presents these contexts to the physician. By gradually adding new facts to the user description the physician fine-tunes the recommendations.

The experiment clarified how a statement on an individual patient's situation is going to be matched with the context instances. Matching results in a recommendation list comprising health condition contexts that come close to the patient's context expressed in the physician's statement. This process selects candidate contexts according to a number of rules:

It takes first all terms in the patient statement that relate to body parts or phenomena without considering minimal semantics that represent qualities.

Recommended contexts must address all body parts and (absent or present) phenomena observed by the physician and stated in the examination report. Minimal semantics are used to fine-tune the selection. Nevertheless they influence the ranking of the individual context description. If the person description mentions a "subcentimetric lump" at certain location the engine relegates a description containing a lump with more than one centimeter in diameter to a lower rank.

It is important to note that the engine takes negation as a quality of a concept item. If, for instance, the physician's observation indicates "pleural effusion" the system no longer considers a context descriptions that explicitly excludes this symptom in its observation section.

Let us consider, for example, a situation in radiology: The physician observes "bronchial dilation in the lower lobes of the left lung". The pattern recognition engine transforms this statement into a patient context description as shown in Figure 5.

Based on the body parts and phenomena mentioned, the context engine finds 27 candidate descriptions. Taking into account the location quality the engine reduces the number of candidates to 13. If the physician adds just the expression “secretion” the list of recommendation contains not more than 4 candidate descriptions.

```
<observations>
    <lungs>
        <dilatation>
            <location>left, bottom</location>
        </dilatation>
    </lungs>
</observations>
```

Fig. 5. A patient context description (“bronchial dilation in the lower lobes of the left lung”) expressed in CDL

We acknowledge that we get such a favorable result because of the relatively small set of candidate contexts. Nevertheless, our database comprises reports on respiratory health problems recorded over a period of about six months by the radiology department. Even if we take into account the possibility that this collection grows to a size 10-fold of the one of our example base, we will see the number of recommendations growing just to 40 cases. A pre-requisite, however, to keep the size of the recommendation list actionable, would be the classification of each patient description along the main health problems before the pattern recognition engine starts its work.

As an initial proof of concept, our work was validated by the user group of physicians involved. It showed us that our stepwise analysis of text patterns yields a controlled vocabulary that, classified along the generic concepts mentioned above, can provide machine-processable descriptions of health contexts. In a next step, we are going to expand this experiment to further health problems including more medical professionals in order to cover also a broader variety of first-time statements on patients’ health conditions.

5 Conclusions

Mobile Computing sees many initiatives, projects, and product developments under way that monitor key indicators of a person’s health situation. They all address portable devices that act as sensors gathering health data such as pulse, temperature and so on besides the coordinates of the person’s physical location. Connected to a remote data mining system, the user of such a wearable device can expect to get alerts warning her that her health situation sparks concerns. Or she may be warned when she is entering an area where other persons experienced respiratory problems.

The context constructed in these location-based services is exclusively built on numerical information and the corresponding mathematical models. This approach is beneficial for a huge number of mobile and embedded applications that operate on sensory data.

With our paper we want to broaden the focus on the benefits of mining textual information for the design of context-awareness. Many medical facts are expressed in natural language statements. Fortunately, the language used in reports describing health conditions often lacks the ambiguity that usually can be found in natural language texts. Due to this quality medical reports lend themselves for the development of a meta-language that identifies the meaning of terms in the light of depicting the context of a patient's health situation. This meta-language defines a domain-dependent Context Description Language that controls text pattern recognition engines and context engines.

By broadening the focus on textual facts, i.e. facts extracted from unstructured texts, we can contribute to the informative performance of context-aware systems. Our context- and ontology-free approach is not restricted to health-related topics. However, it thematically scales only when we consider domains where facts are represented in the same clear manner. The ambiguity rises in textual information the more we need ontologies for a clear identification of meaning in text.

References

1. Mehra, P.: Context-Aware Computing. Beyond Search and Location-Based Services. IEEE Internet Computing 16(2), 12–16 (2012)
2. Fröhlich, P., Oulasvirta, A., Baldauf, Nurminen, A.: On the Move, Wirelessly Connected to the World. Communications of the ACM 54(1), 132–138 (2011)
3. McKenna, M.: The New Age of Medical Monitoring. Scientific American 308(3), 16–17 (2013)
4. Goodchild, M.: Citizens as sensors: The world of volunteered geography. GeoJournal 69, 211–221 (2007)
5. Mehra, P.: Context-Aware Computing. Beyond Search and Location-Based Services. IEEE Internet Computing 16(2), 12–16 (2012)
6. Lee, J.-H., Kim, J.-T., Lee, H.-K., Paik, E.-H.: Design and implementation of the Geo-Context Engine for semantic social media service. In: Conference for Internet Technology and Secured Transactions (ICITST), pp. 383–387 (2011)
7. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax (2004), http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
8. Horrocks, I.: Ontologies and the Semantic Web. Communications of the ACM 51(12), 67 (2008)
9. McGuiness, D.L., van Harmelen, F.: OWL Web Ontology Language (2004), http://www.w3c.org//TR/owl-features/
10. Mernik, M., Heering, J., Sloane, A.M.: When and How Develop Domain-Specific Languages. ACM Computing Surveys 37(4), 316–344 (2005)
11. Stahl, T., Voelter, M.: Model-Driven Software Development. Wiley & Sons (2006)
12. Cao, L., Ramesh, B., Rossi, M.: Are Domain-Specific Models Easier to Maintain Than UML Models? IEEE Software 26(4), 19–21 (2009)
13. Iwanska, L.M.: Natural Language Is a Powerful Knowledge Representation System: The UNO Model. In: Iwanska, L.M., Shapiro, S.C. (eds.) Natural Language Processing and Knowledge Representation, pp. 7–64. AAAI Press, Menlo Park (2000)

14. Agosti, M., Gradegnio, G., Marchetti, P.: A hypertext environment for interacting with large databases. Information Processing and Management 28, 371–387 (1992)
15. Englmeier, K., Koinig, R.: Domain-Specific Deployment and Configuration Language for Composition and Adaptation of Coarse-Grained Services. In: Proceedings of IEEE-SCC 2009 conference, pp. 490–493 (2009)
16. WHO: International Classification of Diseases (ICD) (2013), http://www.who.int/classifications/icd/en/# (retrieved on February 20, 2013)
17. Mahmud, U., Mohammed, Y.J.: Context Inference Engine (CiE): Inferring Context. International Journal of Advanced Pervasive and Ubiquitous Computing 4(3), 13–41 (2012)
18. Garcia-Molina, H., Koutrika, G., Parameswaran, A.: Information seeking: convergence of search, recommendations, and advertising. Communications of the ACM 54(11), 121–130 (2011)

Eccentricity-Based Data Gathering and Diameter-Based Data Forwarding in 3D Wireless Sensor Networks

A.S.M. Sanwar Hosen and Gi-hwan Cho*

Div. of CSE (Cloud Open R&D Center), Chonbuk University, Jeonju, S. Korea
{sanwar,ghcho}@chonbuk.ac.kr

Abstract. This paper proposes an efficient method of constructing three dimensional (3D) wireless sensor networks (WSNs) with aiming to minimization of the overall routing cost. It tries to divide the network into subspaces, and elects a routing centroid node in the eccentricity region from any node in a subspace in terms of minimizing communication cost of that space. The node in an eccentricity region is naturally close to the distance of the radius of that subspace. As a result, the centroid node can forward the gathered data to the node on the diameter. The minimization of the path cost in data gathering and forwarding towards the sink is an efficient approach to design a cost effective 3D WSNs.

Keywords: Wireless Sensor Network, 3D space, Routing Cost, Eccentricity, Diameter, Radius.

1 Introduction

Two dimensional WSNs are generally designed with an assumption that all sensor nodes of a network are deployed on a plane. It may no longer be valid if a network is deployed in the Space, the atmosphere, or the ocean, where sensor nodes need to be distributed over a 3D space. A 3D network is a geometric tri-parametric (length, width, and depth/height) model of the physical environment in which the sensor nodes can be deployed. For example, weather forecasting, climate monitoring, and ocean column monitoring require the sensor nodes to be placed at different heights of the atmosphere and different depths of the sea, thus creating 3D WSNs [1].

Considering the several constraints of a sensor device, it is challenging task to design an efficient WSN in 3D space for which it will be feasible to resolve the issues corresponding to the fundamental problems, such as:

(i) What is the optimal way to deploy the sensor nodes in a 3D space such that the requirement for a minimized number of necessary nodes?

(ii) What is an efficient solution in order to aggregation and routing data to multiple sinks with minimum routing cost?

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_65,

Research towards the above answer of the questions has both theoretical and practical significance in terms of feasibility of 3D WSNs design as well as making long the WSNs' lifetime. Firstly, minimizing the number of deployed nodes to achieve full-coverage and connectivity is very important because the sensor nodes deployed in a 3D space are comparatively expensive [2]. Secondly, making long the lifetime of WSNs in a 3D space requires an effective data aggregation and routing techniques in terms of routing cost minimization.

In this paper, we propose an efficient approach for the data gathering at the eccentricity region of a node in a 3D subspace, along with data forwarding using a node on the diameter of that 3D subspace in order to reduce the hops towards the sinks. For this, we first suggest to partition the entire network into subspaces. Then, the routing centroid node selection in the eccentric region of a subspace is naturally close to the average radius of that subspace. As a result, all deployed nodes in a 3D subspace can communicate with the routing centroid node which can forward the gathered data to the node on the diameter. This approach results in minimizing the path cost to the sink, along with benefits of effectively gathering, merging and forwarding the data.

2 Related Works

In recent years, 3D networks have gained attention in WSNs. To design an efficient network in 2D and 3D space, the most important task is to solve out the coverage with full-connectivity and the routing cost problem. The coverage with full-connectivity problem studied in several works [3]. Some works concentrate on network construction methods, with emphasizing energy consumption. [4] delivers two positions based routing algorithm which is based on the restricted directional floating-based routing mechanism for 3D space. [5] proposes a GDSTR-3D routing algorithm which with 3D convex hulls and spanning tree strategies it forwards packets greedily as long as it could discover a neighbor closer to the destination than the current node. [6] originally presents an energy efficient localized 3D greedy routing algorithm (ERGrd) in WSNs that is a simple variant of greedy algorithm in 3D random space. Henceforth, [7] gives a more expand ERGrd in large-scale random WSNs, which limits its forwarding path inside a restricted 3D convex region.

However, all aforementioned routing algorithms are either based on partial or a full-flooding for data and control packets which result to higher routing cost and incur higher communication overhead. Our method differs from the related works in terms of data gathering in a routing centroid node instead of the full-flooding that intends to minimize the internal communication cost within a subspace. Meanwhile, this approach encourages selecting a forwarding node on the diameter of a 3D subspace to route data with minimized path cost compared to the flooding in a 3D network.

3 EGDF: Eccentricity-Based Data Gathering and Diameter-Based Data Forwarding

This paper introduces a 3D network model inspired from the spanning trees (STs) and optimization problems [8]. While a spanning tree spans all nodes of a given sensor deployed 3D space, a Steiner tree spans a given subset of sensor nodes. Our method proposes that the network can be partitioned into Steiner subsets of sensor nodes, referred to as 3D subspace. In each subspace, a node which is in the eccentricity region and minimizes the internal routing cost should be the data gathering node. The gathered data need to be forwarded towards the sink along the minimized routing cost path to reduce the overall routing cost of the network.

3.1 Routing Cost Estimation on Connected 3D Subspaces

Firstly, we introduce the terms; T is the tree which contains deployed sensor nodes in a 3D network, $l(T,e)$ is the routing load on each communication link e between two communicating nodes, V is the set of all deployed nodes in a 3D space, $E(T)$ is the set of all possible combinations of any source to any destination with bi-directional links, SP_T is a spanning tree considered with shortest path in-between the 3D subspaces that contains minimum number of nodes, w is the assigned weights based on the distance on each link, and the $Cost(T)$ is the overall routing cost of a particular 3D network. The routing cost of a tree is the sum of overall routing cost for the pair nodes in the tree $Cost(T) = \sum_{u,v} d_T(u,v)$, where $d_T(u,v)$ is the distance between u and v on T.

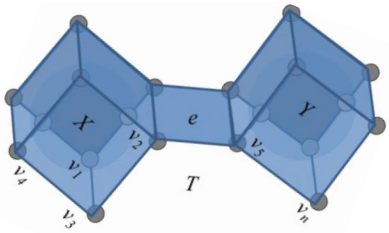


Fig. 1. An example of a T of connected two 3D subspaces

Definition 1. Let T be a tree and $e \in E(T)$. Assume X and Y are the two 3D subspaces which are resulting in removing e from T. The routing load on edge e is defined by $l(T, e) = 2|V(X)||V(Y)|$.

Lemma 1. For a tree T with link weight (distance) w, the overall routing cost is defined as;

$$Cost(T) = \sum_{e \in E(T)} l(T,e)w(e)$$

Proof. Let SP_T denotes a simple path between nodes u and v on the tree T. Then $Cost(T)$ can be obtained as;

$$
\begin{aligned}
Cost(T) &= \sum_{u,v \in V(T)} \left(\sum_{e \in SP_T(u,v)} w(e) \right) \\
&= \sum_{e \in E(T)} \left(\sum_{u \in V(T)} |\{v \mid e \in SP_T(u,v)\}| \right) w(e) \\
&= \sum_{e \in E(T)} l(T,e) w(e)
\end{aligned} \tag{1}
$$

3.2 Routing Centroid Node Selection in the Eccentric Region of a 3D Subspace

To generalize our idea, we propose a 3D network model, in which the 3D space is divided into 3D subspaces. In each 3D subspace, a node would be the routing centroid node among the constituted nodes. Our method intends to minimize the routing cost of intercommunication among the constituent nodes within a particular 3D subspace. Therefore, among the nodes located within the eccentricity region in a subspace, the node having the minimum routing cost would be the routing centroid node.

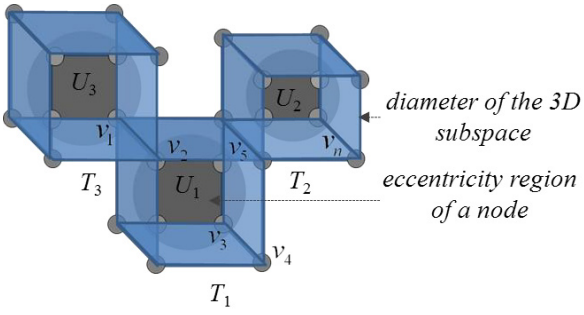


Fig. 2. An example of a node deployed in a 3D cube shape connected space

Let $G = (V, E, w)$ be a 3D space and $U \subset V$, where $U = \{U_1, U_2, U_3, \ldots, U_k\}$ obtain the deployed nodes $\{v_1, v_2, v_3, \ldots, v_n\} \in V$ in a 3D subspace. By $D_G(v,U)$ where $v \in U$, we denote the maximum distance from a node to any other node in a 3D subspace. For a node v, the eccentricity of v is the maximum of the distance to any other node in that subspace, i.e., $max_u \in V\{d_G(u,v)$ or $D_G(v,V)$. Therefore, the routing centroid node in the eccentricity region of any node in that 3D subspace can communicate to any node using a maximum transmission distance that is not more than the radius of the subspace. For example, the nodes $\{(v_2, v_5, v_3), (v_1, v_2, v_3)\} \in U_1$ is in the T_1 of the cube and the octahedral 3D shape subspaces are shown in Fig. 2 and 3. respectively.

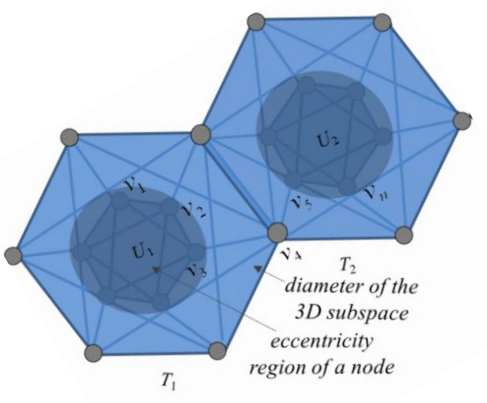


Fig. 3. An example of a node deployed in a 3D octahedral shape connected space

The node $\{v_i \in U_i\} \in T_i$ which obtains the minimum internal communication cost is selected as the routing centroid node among the constituent nodes of the 3D subspace. This is the node that obtains minimum internal routing cost according to Eq. (2).

$$Cost(T_i) = \min \sum_{e \in E(T_i)} l(T_i, e) w(e) \tag{2}$$

3.3 Data Forwarding Node Selection on Diameter of the 3D Subspace

We aim to gather data at the routing centroid node from constituent sensor nodes, and to forward it to the sink(s) through the minimum routing cost path using the nodes on the diameter of a 3D subspace as intermediate nodes.

The diameter of a 3D subspace is the longest distance between any two nodes in a 3D subspace. And the radius of a 3D subspace is the minimum eccentricity among all nodes in a 3D subspace and a center of a 3D subspace is a node with eccentricity equal to the radius. Therefore, our method emphasizes to elect a centroid node in the eccentricity region of a 3D subspace which is not more than the distance of its radius. So, any node in a 3D subspace can communicate using minimum distance, where 2×radius-1≤ diameter. This communicating path $P = \{P_1, P_2,...,P_n\}$ can be divided into three sub-paths: the path i belong in a terminal node (source) to a routing centroid node $v_i \in U_i$, the path containing nonterminal node(s) (intermediate node) on the diameter of a 3D subspace form routing centroid node, and the path j belongs a routing centroid node to any sink(s). For convenience, we can define

$$Cost(P_i) = \min \sum_{e \in E(P_i)} l(P_i, e) w(e)$$

$$d_T(u, v) \le d_T(v, P) + d_T^P(u, v) + d_T(u, P) \tag{3}$$

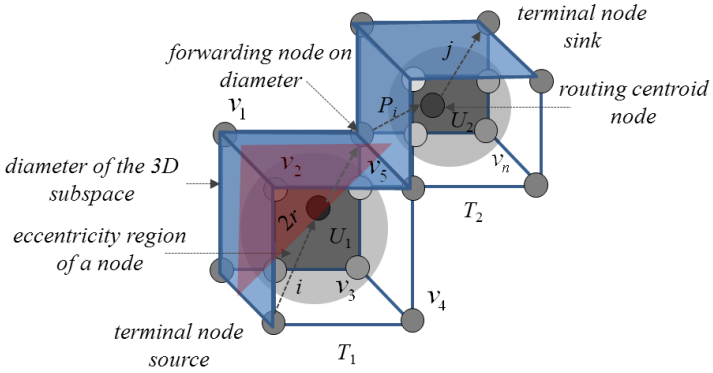


Fig. 4. An example of a data forwarding path in-between two subspaces

In the data routing, a routing centroid node can select the path which obtains the minimum routing cost as below

$$Cost(P_i) = \min \sum_{e \in E(P_i)} l(P_i, e) w(e) \tag{4}$$

Therefore, the total cost of the tree T would minimize in the entire network as in the following equation, where k is the number of connected 3D subspaces and n is for number of forwarding paths in the entire network.

$$Cost(T) = \min \left(\sum_{i=1}^{k} l(T_i, e) w(e) + \sum_{i=1}^{n} l(P_i, e) w(e) \right) \tag{5}$$

4 Conclusion

We tried to minimize the intercommunication cost in 3D subspaces and select a minimum cost path between the different subspaces in the routing domain. The routing centroid node in the eccentricity region of a subspace is convenient to communicate with each one of the other nodes in that particular space. Selecting a node on the diameter for data forwarding is an efficient approach by maximizing the transmission range of forwarding nodes. The node selection on the diameter of subspace for data forwarding is competent to reducing the number of intermediate nodes. So, we believe that our method plays an important role in the designing of routing mechanism in 3D WSNs. As a future plan, we would like to demonstrate this method in practice.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(KRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2042035).

References

1. Akyildiz, I., Pompili, D., Melodia, T.: Underwater Acoustic Sensor Networks: Research Challenges. Ad Hoc Networks Journal (2005)
2. Li, J., Yu, M.: Sensor Coverage in Wireless Ad Hoc Sensor Networks. International Journal of Sensor Networks 2, 218–229 (2007)
3. Bai, X., Zhang, C., Xuan, D., Jia, W.: Full-Coverage and Connectivity, Three Dimensional Networks. In: IEEE INFOCOM, pp. 388–396 (2009)
4. Abdallah, A., Fevens, T., Opatrny, J.: Hybrid Position-Based 3D Routing Algorithms with Partial Flooding. In: ECE, pp. 227–230 (2006)
5. Zhou, J., Chen, Y., Leong, B., Sundaramoorthy, P.: Practical 3D Geographic Routing for Wireless Sensor Networks. In: 18th ACM ENSS, pp. 337–350 (2010)
6. Huang, M., Li, F., Wang, Y.: Energy-Efficient Restricted Greedy Routing for Three Dimensional Random Wireless Networks. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 95–104. Springer, Heidelberg (2010)
7. Wang, Y., Yi, C., Huang, M., Li, F.: Three-Dimensional Greedy Routing in Large-Scale Random Wireless Sensor Networks. Ad Hoc Networks 11(4) (2013)
8. Wu, B., Chao, K.: Spanning Trees and Optimization Problems. Chapman&Hall (2004)

Weighted Mining Frequent Itemsets Using FP-Tree Based on RFM for Personalized u-Commerce Recommendation System

Young Sung Cho[1] and Song Chul Moon[2]

[1] Department of Computer Science, Chungbuk National University, Cheongju, Korea
[2] Department of Computer Science, Namseoul University, Cheonan-city, Korea, Korea
youngscho@empal.com, moon@nsu.ac.kr

Abstract. This paper proposes a new weighted mining frequent itemsets using FP-tree based on RFM for personalized u-commerce recommendation system under ubiquitous computing. Existing recommendation system using association rules still does not only reflect exact attributes of item but also has the problem, such as delay of processing speed from a cause of frequent scanning a large data, scalability and accuracy. In this paper, to solve these problems, it is necessary for us to make RFM(Recency, Frequency, Monetary) score of item and to extract the most frequently purchased data from the whole data in order to improve the accuracy of recommendation, to consider frequently changing the weighted patterns by emphasizing the important items with high purchasability according to the threshold for creative the weighted mining frequent itemsets using FP-tree without occurrence of candidate set. To verify improved performance, we make experiments with dataset collected in a cosmetic internet shopping mall.

Keywords: Association Rules, RFM, Weighted Mining Frequent Itemsets using FP-tree.

1 Introduction

Along with the advent of ubiquitous computing environment and the spread of intelligent portable device such as smart phone, PDA and smart pad has been amplified, a variety of services and the amount of information has also increased. It is becoming a part of our common life style that the demands for enjoying the wireless internet are increasing anytime or anyplace without any restriction of time and place[1],[4]. The customers need a recommendation system that can recommend item which they really want on behalf of them. In the u-commerce, it is important to recommend the proper item among large item sets. Therefore, if the recommendation system can recommend the suitable item which they really want, the customers are satisfied with the system. The possession of intelligent recommendation system is becoming the company's business strategy. A personalized recommendation system using data mining technique based on RFM to meet the needs of customers has been actually processed the research[1-5]. We can improve the accuracy of recommendation through an weighted mining frequent itemsets using FP-tree without

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_66,

occurence of candidate set so as to be able to generate the associated items' rules. As a result, we propose a new weighted mining frequent itemsets using FP-tree based on the most frequently purchased data extracted from the whole data for recommendation in u-commerce under ubiquitous computing environment. The next Sect. briefly reviews the literature related to studies. The Sect. 3 is described a new method for personalized recommendation system in detail, such as system architecture with sub modules, the procedure of processing the recommendation, the algorithm for proposing method. The Sect. 4 describes the evaluation of this system in order to prove the criteria of logicality and efficiency through the implementation and the experiment. In Sect. 5, finally it is described the conclusion of paper and further research direction.

2 Related Works

2.1 RFM

RFM segments customers on the basis of how long since they made purchases, how frequently they make purchases, and how much money they spend for segmentation by product usage. The RFM score can be a basis factor how to determine purchasing behavior on the internet shopping mall, is helpful to buy the item which they really want by the personalized recommendation. One well-known commercial approach uses five bins per attributes, which yields 125 cells of segment. The following expression presents RFM score to be able to create an RFM analysis. The RFM score will be shown how to determine the customer as follows, will be used in this paper. The variables (A, B, C) are weights. The categories (R, F, M) have five bins.

$$\text{RFM score} = A \times R \ + \ B \times F \ + \ C \times M \tag{1}$$

The RFM score is correlated to the interest of e-commerce[2]. It is necessary for us to keep the analysis of RFM to be able to reflect the attributes of the item in order to find the items with high purchasability. In this paper, we can use the customers' data and purchased data with 60.98% in the rate of portion for the purchasing counts[4].

2.2 Association Rules

Association rule mining search for interesting relationship among items in a given database. Association rules, first introduced by Agrawal[6], are frequently used by market basket analysis including cross marketing, recommendation system in e-commerce. Association rules which satisfy a minimum confidence threshold are then generated from the frequent itemsets. The traditional association rule mining employs the support measure, which treats every transaction equally. However, in our real world data sets, the weight importance of a pattern may vary frequently due to some unavoidable situations. Usually in an association rules, it is expressed in the form of the rule $X \rightarrow Y$. The rule of $X \rightarrow Y$ means that the transaction including the item of X tends to include the itemsets. And then in an weighted association rules, the w-support of an weighted association rule $X \rightarrow Y$ is defined as

$$\textbf{WSUPP}(\mathbf{X} \rightarrow \mathbf{Y}) = WSUPP(X \cup Y) \quad (2)$$

and the w-confidence is

$$\text{WCONF}(X \rightarrow Y) = \frac{WSUPP(X \cup Y)}{WSUPP(X)} \quad (3)$$

Basically, w-support measures how significantly X and Y appear together; w-confidence measures how strong the rule is. An weighted association rule mining becomes an important research issue in data mining and knowledge discovery by considering different weights for different items. It is necessary to consider these dynamic changes in different application area such as retail market basket data analysis. Much effort has been dedicated to association rule mining with pre-assigned weights[8],[9]. It is crucial to have different weights for different transactions in order to reflect their different importance and adjust the mining results by emphasizing the important transactions.

2.3 Mining Frequent Itemsets Using FP-Tree

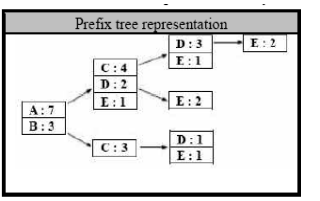
Han et al. [10] proposed a data structure called the FP-tree(Frequent Pattern tree). The FP-tree is a compact representation of all relevant frequency information in a database. Every branch of the FP-tree represents a frequent itemset and the nodes along the branches are stored in decreasing order of frequency of the corresponding items with leaves representing the least frequent items. Compression is achieved by building the tree in such a way that overlapping itemsets share prefixes of the corresponding branches[11]. The FP-tree T has a header table, T:header, associated with it. Single items and their counts are stored in the header table in decreasing order of their frequency. The entry for an item also contains the head of a list that links all the corresponding nodes of the FP-tree. Compared with breadth-first algorithms such as Apriori and its variants, which may need as many database scans as the length of the longest pattern, the FP-growth method only needs two database scans when mining all frequent itemsets. The first scan counts the number of occurrences of each item. The second scan constructs the initial FP-tree which contains all frequency information of the original dataset. Mining the database then becomes mining the FP-tree. To construct the FP-tree, the first scan is to find all frequent items by an initial scan of the database. Then, these items are inserted into the header table in decreasing order of their count. In the next scan, as each transaction is scanned, the set of frequent items in it are inserted into the FP-tree as a branch[11]. If an itemset shares a prefix with an itemset already in the tree, the new itemset will share a prefix of the branch representing that itemset. In addition, a counter is associated with each node in the tree. The counter stores the number of transactions containing the itemset represented by the path from the root to the node in question. This counter is updated during the second scan, when a transaction causes the insertion of a new branch. In this paper, we can use the weighted mining frequent itemsets using FP-tree based on RFM(WMFP) for generating the weighted association rules. In this paper, we can use the weighted mining frequent itemsets using FP-tree based on the customers' data and

Table 1. Example transaction database involving items A,B,C,D, and E

Transaction DB	Lexicographically sorted
{A, D, E}	{A, C, D}
{B, C, D}	{A, C, D, E}
{A, C, E}	{A, C, D, E}
{A, C, D, E}	{A, C, E}
{A, E}	{A, D, E}

Transaction DB	Lexicographically sorted
{A, C, D}	{A, D, E}
{B, C}	{A, E}
{A, C, D, E}	{B, C}
{B, C, E}	{B, C, D}
{A, D, E}	{B, C, E}

Table 2. Prefix tree representation by FP-tree mining



the most frequently purchased data extracted from the whole data to recommend the item they really want exactly.

3 Our Proposal for a Personalized u-Commerce Recommendation System

3.1 System Architecture

We can depict the system configuration concerning the personalized u-commerce recommendation system using the weighted mining frequent itemsets using FP-tree based on RFM under ubiquitous computing environment. This system had four agent modules which have the analytical agent, the recommendation agent, the learning agent, the data mining agent in the internet shopping mall environment. We observed the web standard in the web development, so developed the interface of internet to use full browsing in mobile device. As a matter of course, we can use web browser in wired internet to use our recommendation system. We can use the system under WAP in mobile web environment using feature phone as well as using the internet browser such as safari browser of iPhone and Google chrome browser based on android so as to use our system by using smart phone.

3.2 Weighted Mining Frequent Itemsets Using FP-Tree Based on RFM for Personalized u-Commerce Recommendation System

In this part, we can depict the weighted mining frequent itemsets using FP-tree based on RFM for personalized u-commerce recommendation system. Our algorithm can consider situation where the weight / importance of a pattern may vary dynamically in e-commerce on the real world. It is necessary for us to consider the quantity of purchased data extracted by the scope of RFM score which is between the score is more than 19 points and the score is less than 40 points, had a lot of purchasing counts in order to prevent delay of processing speed from a cause of frequent scanning a large data. We can calculate the rate of weight based on the quantity item by each rank of the RFM score. We can have different weights for different transactions and to generate the weighted association rules through the weighted mining frequent itemsets using FP-tree based on RFM. Thus we can forecast frequently changing trends by emphasizing

the important items. At first, we can aggregate the quantity of purchased data(sale_dat3) extracted from the whole data (sale). After that, we can make the rate of weight using aggregated counts of a section, that is, it is become the value of weight based on the quantity of extracted data. To propose our paper, we have two step of procedural preprocessing. The 1st step is that it is necessary for us to make the RFM scores to reflect the attributes of the item. The 2nd step is that it is necessary for us to process an weighted mining frequent itemsets using FP-tree based on the most frequently purchased data extracted from the whole data in order to forecast frequently changing trends by emphasizing the important items with high purchasability according to the threshold for creative the weighted association rules.for generating association rules with w-support, w-confidence and w-lift through the weighted mining frequent itemsets using FP-tree based on RFM(WMFP). The procedural algorithm of WMFP for personalized u-commerce recommendation system is depicted as the following Table 3.

Table 3. The procedural algorithm of WMFP for personalized u-commerce recommendation system

Step 1 : The RFM score of customer is computed so as to reflect the attributes of the customer, consists of three attributes(R, F, M), each attribute has five bins divided by each 20%, exact quintile.
Step 2 : The system can aggregate the quantity of purchased data by each interval customer's RFM scores, which is aggregated counts of distribution from the whole data, make the rate of weight.
Step 3 : The system can calculate the rate of weight based on quantity item with each rank of RFM score for customer.
Step 4 : The system can scan extracted database(sale_dat3) and make the weighted mining frequent itemsets using FP-tree based on sale_dat3.
Step 5 : Weighted Association rules are created by WMFP
Step 6 : Wsupport = $\sum^{n}$ *Weight/N X Support count /* N is numbers of item in the rules */*
Step 7 : The system *can create creative the weighted association rules with w- support, w-confidence and w-lift through the weighted mining frequent itemsets using FP-tree using FP-tree without occurrence of candidate set.*
Step 8 : The system can reflect frequently changing the weighted patterns by emphasizing the important items.
Step9: The system can recommend the items with high purchasability according to the threshold for creative the weighted association rules with w-support, w-confidence and w-lift through WMFP.

3.3 The Procedural Algorithm for Personalized u-Commerce Recommendation System

The system can search the information in the cluster selected by using the code of classification and customer's RFM score in users' information. It can scan the preference as the average of brand item in the cluster, suggest the brand item in item category selected by the highest probability for preference as the average of brand item. This system can create the list of recommendation with TOP-N of the highest preference of item to recommend the item with purchasability efficiently. This system can recommend the items with efficiency, are used to generate the recommendable item according to the basic threshold for the weighted mining frequent itemsets using FP-tree, with w-support, w-confidence and w-lift. It can recommend the associated item to TOP-N of recommending list if users want to have the cross-selling or up-selling. This system takes the cross comparison with purchased data in order to avoid the duplicated recommendation which it has ever taken.

4 The Environment of Implementation and Experiment and Evaluation

4.1 Experimental Environment

This system proposes a new weighted mining frequent itemsets using FP-tree based on RFM under ubiquitous computing environment. In order to do that, we make the implementation for prototyping of the internet shopping mall which handles the cosmetics professionally and do the experiment. It is the environment of implementation and experiment below.

- OS: Windows XP SP2,
- Web Server: Apache 2.2.14 / WAP 2.0
- Server-Side Script : JSP/PHP 5.2.12
- XML/WML2.0/ HTML5.0/CSS3/JAVASCRIPT
- Database: MySQL 5.1.39
- J2SDK(1.7.0_11)
- MySQL JDBC
- J2SDK(1.7.0_11)
- jakarta-tomcat (5.0.28)

We have carried out the implementation and the experiment for proposing system through system design, we have finished the system implementation about prototyping recommendation system. It could be improved and evaluated to new system through the result of experiment with the metrics such as precision, recall, F-measure as comparing proposing ystem(W_FP) with other previous system(O_ARM) using original method of mining and existing system.

4.2 Experimental Data for Evaluation

We used 319 users who have had the experience to buy items in e-shopping mall, 580 cosmetic items used in current industry, 1600 results of purchase data recommended in order to evaluate the proposing system[4]. In order to do that, we make the implementation for prototyping of the internet shopping mall which handles the cosmetics professionally and do the experiment. We have finished the system implementation about prototyping recommendation system. We'd try to carry out the experiments in the same condition with dataset collected in a cosmetic internet shopping mall. It could be evaluated in MAE and Precision, Recall, F-measure for the recommendation system in clusters. It could be proved by the experiment through the experiment with learning data set for 12 months, testing data set for 3 months in a cosmetic cyber shopping mall[4]. The 1st system of the weighted mining association rules based quantity item with RFM score, is proposing method(W_FP) called by "proposal", the 2nd system is the original method(O_ARM) using the ordinary association rules mining, the third system is existing system. The proposing method's overall performance evaluation for recommendation is precision, recall and F-measure as comparing proposing method using (W_FP) and the original method using (O_ARM). The performance was performed to prove the validity of recommendation and the system's overall performance evaluation. The metrics of evaluation for recommendation system in our system was used in the field of information retrieval commonly[12].

Table 4. The result for table of precision, recall, F-measure for recommendation ratio by each cluster

Cluster	Proposal(W_FP)			O_ARM			Existing		
	Precision1	Recall1	F-measure1	Precision2	Recall2	F-measure2	Precision3	Recall3	F-measure3
C1	42.23	80.66	55.44	46.20	55.70	45.90	48.79	31.32	35.64
C2	45.25	84.47	58.93	43.20	52.53	45.76	49.36	29.54	35.06
C3	46.42	76.94	57.90	40.99	23.93	30.05	44.26	21.81	27.65
C4	42.81	64.52	51.47	56.06	38.37	43.77	52.49	34.98	39.75
C5	54.99	78.45	64.66	44.83	47.40	42.74	50.41	43.21	43.10
C6	48.56	81.40	60.83	33.56	37.23	34.68	50.93	36.60	39.64
C7	41.58	74.08	53.26	53.94	27.27	34.90	47.41	26.81	32.26
C8	53.92	78.87	64.05	45.07	37.23	38.29	43.60	36.60	37.82
C9	27.72	74.08	40.34	64.08	28.45	37.19	46.68	25.19	30.28
C10	59.99	93.78	73.17	73.85	62.50	64.42	67.23	55.34	57.10

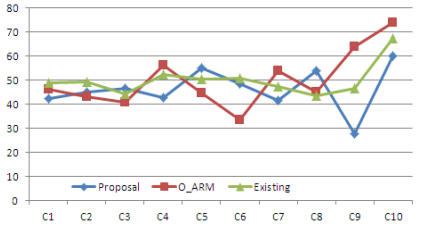


Fig. 1. The result of recommending ratio by precision

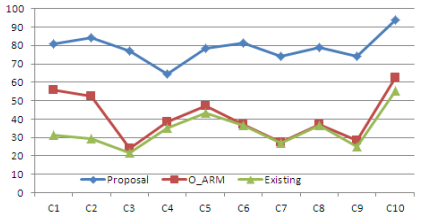


Fig. 2. The result of recommending ratio by recall

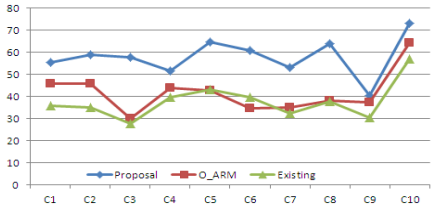


Fig. 3. The result of recommending ratio by F-measure

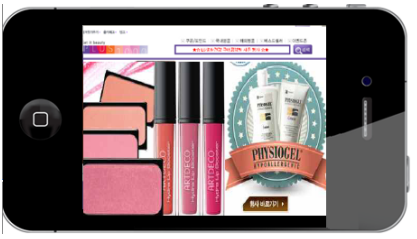


Fig. 4. The result of recommending items of cosmetics

Above Table 4 presents the result of evaluation metrics (precision, recall and F-measure) for recommendation system. The weighted mining frequent itemsets using FP-tree is improved better performance of proposing method using (W_FP) than the original method using (O_ARM). The proposed higher 37.66% in recall even if it is lower 3.83% in precision than the original method using (O_ARM), higher 16.24% in F-measure than the original method. After that, it shows that our algorithm is very efficient and scalable for the recommendation system. Above figure 4 is shown in the result of screen on a smart phone. The performance of proposing mining method was improved more counts of support and rule than the original method, it was especially worthy of notice, in the rule counts, had an effect about 4 times what the original mining method did before. As a result, it was efficient for us to recommend the items of association because it is strong cohesion of the attribute of item based the weighted association rules using the weight based quantity item with RFM score. So, we could have the recommendation system to be able to recommend the items with high purchasability. Above figure 4 is shown in the result of screen on a smart phone. The performance of proposing method is improved although it is less in average of confidence (average confi_rate), however it is efficient for us to recommend the items of association because it is strong cohesion of the attribute of item because of using a new weighted mining frequent itemsets using FP-tree based on RFM.

5 Conclusion

Recently u-commerce as a application field under ubiquitous computing environment, is in the limelight[5]. Existing recommendation system using association rules still does not only reflect exact attributes of item but also has the problem, such as delay of processing speed from a cause of frequent scanning a large data, scalability and accuracy. And also, existing algorithms for mining association rules are based on fixed weight, do not reflect the weight / importance of a pattern, and do not consider these dynamic changes in different application area such as retail market basket data analysis. It was necessary for us to make RFM score of item and to extract the most frequently purchased data from the whole data in order to improve the accuracy of recommendation, to reflect frequently changing the weighted patterns by emphasizing the important items with high purchasability according to the threshold for creative the weighted mining frequent itemsets using FP-tree without occurrence of candidate set in order to solve these problems, to use the dynamic weights in proposing method of mining. We could improve the performance of the weighted mining rapidly. As a result, we proposed a new weighted mining technique using FP-tree for personalized u-commerce recommendation system in real datasets environment in order to improve the accuracy of recommendation with high purchasability. As a matter of course, we have described that the performance of the proposing method with the weighted mining using FP-tree based RFM was improved better than the original method and existing system in mining test. It is meaningful to present a new weighted mining technique using FP-tree based on quantity item with RFM score for personalized u-commerce recommendation system under ubiquitous computing environment. The following research will be looking for a personalized recommendation in semantic web environment by fuzzy clustering approach to increase the efficiency and scalability under ubiquitous computing environment**.**

Acknowledgements. This paper[2)] was supported by funding of Namseoul University.

References

1. Cho, Y.S., Ryu, K.H.: Personalized Recommendation System using FP-tree Mining based on RFM. In: KSCI, vol. 17(2) (February 2012)
2. Jin, B.W., Cho, Y.S., Ryu, K.H.: Personalized e-Commerce Recommendation System using RFM method and Association Rules. In: KSCI, vol. 15(12), pp. 227–235 (December 2010)
3. Cho, Y.S., Moon, S.C., Noh, S.C., Ryu, K.H.: Implementation of Personalized recommendation System using k-means Clustering of Item Category based on RFM. In: 2012 IEEE International Conference on Management of Innovation & Technology Publication (June 2012)
4. Cho, Y.S., Moon, S.C., Jeong, S.P., Oh, I.B., Ryu, K.H.: Clustering Method using Item Preference based on RFM for Recommendation System in u-Commerce. In: Han, Y.-H., Park, D.-S., Jia, W., Yeo, S.-S. (eds.) Ubiquitous Information Technologies and Applications. LNEE, vol. 214, pp. 353–362. Springer, Heidelberg (2012)

5. Cho, Y.S., Moon, S.C., Ryu, K.H.: Mining Association Rules using RFM Scoring Method for Personalized u-Commerce Recommendation System in Emerging Data. In: Kim, T.-H., Ramos, C., Abawajy, J., Kang, B.-H., Ślęzak, D., Adeli, H. (eds.) MAS/ASNT 2012. CCIS, vol. 341, pp. 190–198. Springer, Heidelberg (2012)
6. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Datasets. In: Proc. ACM SIGMOD 1993, pp. 207–216 (1993)
7. Agrawal, A., Faloutsos, C., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D.C., pp. 207–216 (May 1993)
8. Ramkumar, G.D., Ranka, S., Tsur, S.: Weighted Association Rules: Model and Algorithm. In: Proc. ACM SIGKDD (1998)
9. Tao, F., Murtagh, F., Farid, M.: Weighted Association Rule Mining Using Weighted Support and Significance Framework. In: Proc. ACM SIGKDD 2003, pp. 661–666 (2003)
10. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2004)
11. Grahne, G., Zhu, J.: Efficiently Using Prefix-trees in Mining Frequent Itemsets. In: Proc. IEEE ICDM 2003 Workshop FIMI 2003. CEUR Workshop Proc. series, vol. 90 (2003), http://ceur-ws.org/vol-90
12. Herlocker, J.L., Kosran, J.A., Borchers, A., Riedl, J.: An Algorithm Framework for Performing Collaborative Filtering. In: Proceedings of the 1999 Conference on Research and Development in Information Research and Development in Information Retrival (1999)

The System of Stress Estimation for the Exposed Gas Pipeline Using the Wireless Tilt Sensor

Jeong Seok Oh[1], Hyo Jung Bng[1], and Si-Hyung Lim[2]

[1] Institute of Gas Safety R&D, Korea Gas Safety Corporation, 11 Soraesan-gil (Daeya-dong), Siheung –Shi, Gyeonggi-do, Korea
[2] School of Mechanical Engineering, Kookmin University, Seoul, Korea
jsoh90@gmail.com, shlim@kookmin.ac.kr

Abstract. Gas pipelines are exposed to the danger especially in bridges, roads and subway construction areas. It can cause leakage accidents from the stress and vibration changes and it can threaten human's life. To avoid that, the gas pipelines should be monitored continuously. The system of stress estimation using MEMS (Micro electro mechanical system_ wireless tilt sensor has been developed and has been evaluated by a lob test bench.

Keywords: MEMS, gas pipeline, tilt sensor, stress, wireless communication.

1 Introduction

In the roads and subway construction sections, pipelines are hung by beams and are exposed for a long time so, quite a number of risk factors are expected to be involved on the construction such as a gas explosion [1, 2]. Some companies are measuring the stress on the pipelines using strain-gage. However, it is hard to utilize the strain-gage on the pipelines. Also, the construction areas are huge to utilize the strain-gage on the pipelines. When the wireless tilt sensor which is based on the MEMS is used, it can be an easy way to estimate the stress on the pipelines [3, 4]. In this study, the system of stress estimation for the exposed pipelines using wireless tilt sensor has been developed and has been tested by a lab test bench which is based on real environment of the construction areas.

2 The Theory of Stress Estimation Using MEMS Tilt Sensor

The stress can be estimated, when the slope is known, equation (1) and (2) are shown the relationship between the stress, moment, and slope [5, 6]. To estimate the stress, the slope is measured my wireless title sensor. The slope is used to make the polynomial equation by curve fitting. Difference between the polynomial equation and the result is multiplied by Young's modulus and moment of inertia and the result is moment. It is possible to estimate the stress by substituting moment in the equation (2).

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_67,

$$M(x) = EI\frac{d\theta}{dx} \tag{1}$$

$$\sigma(x) = \frac{M(x)y_{\max}}{I} \tag{2}$$

Fig. 1 represents the exposed gas pipeline which is hung by a beam. The wireless tilt sensors are installed on the pipeline at regular intervals. There is no necessity for sensors to install at same intervals. However, it will be helpful to equip the sensors with pipelines in a danger zone. When the external force is applied on the pipeline, the pipeline has the deflection which can be measured by wireless tilt sensor. The curve fitting is used to interpolate the slope. When the slope is interpolated, it has to be third-order polynomial. Also, at least four sensors are needed in a section. In the Fig. 1, the concentrated load is applied center along the beam. It causes the same angle with the opposite sign. In this case, it is possible to interpolate as one equation. However, when the concentrated load is got out of the center place, it causes large amount of errors to estimate the stress using one interpolated equation. Therefore, the slope should be interpolated according to angle which has the same sign. If the slope is interpolated by more high order, it could be possible to get the more exact result but it needs more many sensors. Furthermore, there is no big difference between the two. Therefore, third ordered interpolating polynomial has an advantage.

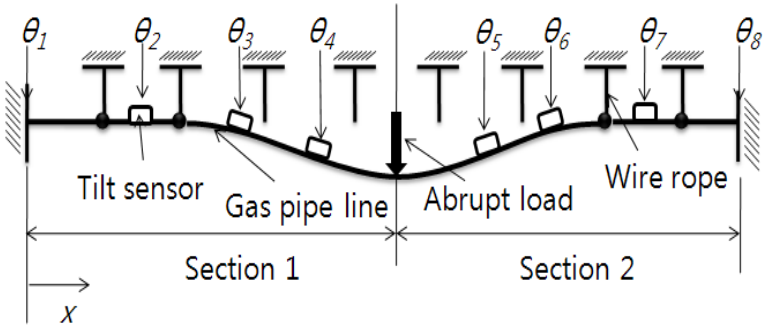


Fig. 1. Schematic of exposed gas pipeline setting with wireless tilt sensor

The method of stress estimation has been investigated by ANSYs which is commercial program of FEM. In the Fig. 2, the pipeline is applied the force of gravity and a concentrated load. The pipeline is fixed at both ends and is applied 98.5 N of concentrated load. There is a node which has different interval to examine that there is no necessity for sensors to install at same intervals. In a section, there are four sensors that have the same sign of angle. The material properties of the pipeline are given by Table 1.

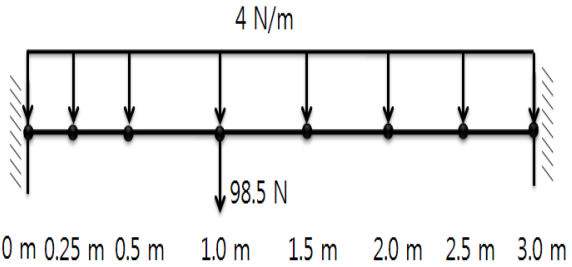


Fig. 2. Distributed and concentrated loads acting on an exposed gas pipeline

Table 1. Material properties of the gas pipeline used in FEM analysis

Outer diameter	Inner diameter	Young's Modulus *(E)*	Density *(ρ)*
0.034 m	0.0275 m	3.5 GPa	1300 kg/m^3

To investigate the method of stress estimation, the Von Mises stress that was obtained from ANSYS has been compared with the stress that is estimated by the method. The ANSYS is used to get the slopes of nodes and then the polynomial is interpolated by third order. Equation (3) and (4) are shown the interpolating polynomials.

$$f(x) = -0.0151x^2 + 0.6012x - 0.3562 \tag{3}$$

$$f'(x) = -0.0152x^2 - 0.1488x + 0.3938 \tag{4}$$

The third equation is interpolated between 0m and 1m; otherwise, the fourth equation is interpolated between 1m and 3m. The interpolating polynomials are differentiated and are substituted in equation (2) to estimate the stress. Stress comparison between the Von Mises stress that is analyzed by ANSYS and the estimated stress as shown Table 2.

Table 2. Stress comparison

Node	Von Mises Stress (Pa)	Stress estimation (Pa)	Error (%)
0m	21.193,000	21,193,900	0.0042
0.25m	12,306,000	12,307,230	0.0010
0.5m	3,523,100	3,532,812	0.0201
1.0m	13,675,000	13,679,050	0,0296
1.5m	8,117,100	8,117,101	0
2.0m	2,105,800	2,105,824	0.0011
2.5m	4,358,500	4,358,486	0.0003
3.0m	11,276,000	11,275,828	0.0015

3 The Testing System and Method

The experimental test bench is mad by aluminum profiler to investigate the method of stress estimation. The test bench is about 3 meters long, 0.25 meters wide, and 1 meter height. To make large slopes, the pipeline is made by aluminum whose material properties are given by Table 3.

Table 3. Material properties of the gas pipeline used in a test bech

Outer diameter	Inner diameter	Young's Modulus *(E)*	Density *(ρ)*
0.016 m	0.014 m	68.9 GPa	2700 kg/m^3

Wireless tilt sensors are attached on the pipeline to measure the slopes. The strain gages are installed because of comparing stress values strain gages and the method of stress estimation.

Fig. 3. MEMS wireless tilt sensor attached on a pipeline of test bench

The pipeline is fixed at both ends and is applied concentrated load of 17.5 N by a weight of 1.8 kg. The estimated stress was compared with measured stress using strain gage and ANSYS. The tilt signals have been measured in 30 Hertz and have been 100 point moving average filter.

4 The Result of the Testing

Fig. 5 represents the stress along the pipeline that is measured in their separate ways. The data was calculated the average of the ten times. Furthermore, Fig. 6 represents stress at each point.

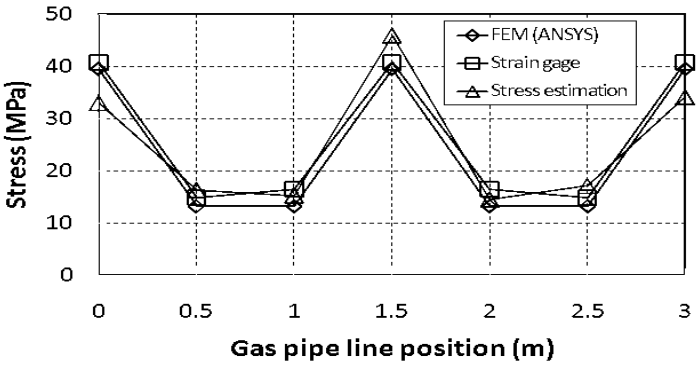


Fig. 4. Stress vs. gas pipeline position

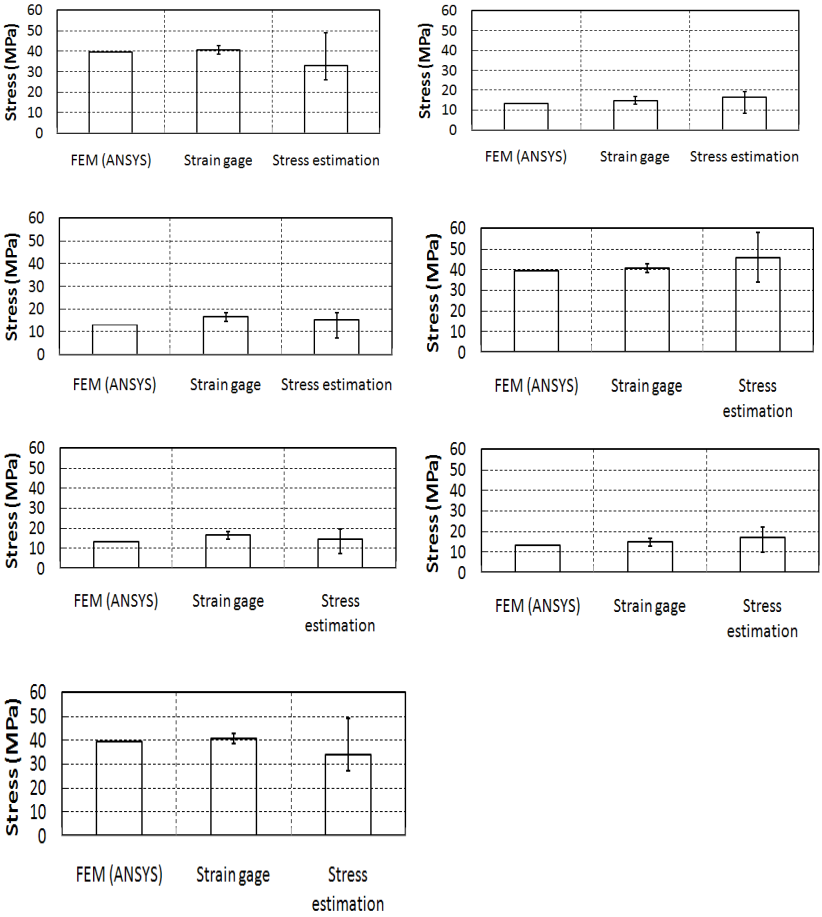


Fig. 5. Stress comparison at various positions (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 m)

On the tests, error rate shows about 13.2% between estimated stress using wireless tilt sensor and measured stress using the strain gages. Noise level of ±0.5° from the wireless tilt sensor is the major cause of the errors.

5 Conclusion

In this paper, the system of stress estimation for the exposed pipelines using wireless tilt sensor has been developed and has been tested in a lab test bench. As a result, the accuracy of the system is measured about 13.2 percentages. If the safety factor is considered and the pipelines are expended, this system will be an easy way to utilize in the construction areas. The stress estimation system is expected to measure the stress of pipelines inexpensively and effectively.

References

1. Hong, S.K.: Development and Application of Exposed Gas Pipeline Monitoring System. Report of Korea Gas Corporation (2003)
2. Oh, J.S., Park, J.S., Kwon, J.R.: Selecting the Wireless Communication Methods for Establishing Ubiquitous City-Gas Facilities in Korea. In: Park, J.H., Chen, H.-H., Atiquzzaman, M., Lee, C., Kim, T.-H., Yeo, S.-S. (eds.) ISA 2009. LNCS, vol. 5576, pp. 823–828. Springer, Heidelberg (2009)
3. Oh, J.S., Bang, H.J., Ko, H.: An Empirical Study on Smart Safety Management Architecture for Gas Facilities in Korea. Information 15(3), 1107–1122 (2011)
4. Jung, H.K.: The best suited Application of Risk Based Inspection. Journal of KIGAS, 83–88 (2008)
5. Shigley, J.E.: Mechanical Engineering Design, pp. 173–212. McGraw-Hill (2005)
6. Hibbeler, R.C.: Mechanics of materials, pp. 569–585. Prentice Hall (2005)

The Architecture Design of Semantic Based Open USN Service Platform Model*

Hyungkyu Lee[2], Namje Park[1,**], and Hyo-Chan Bang[2]

[1] Department of Computer Education, Teachers College, Jeju National University, 61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Korea
namjepark@jejunu.ac.kr
[2] Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, Korea
{leehk,bangs}@etri.re.kr

Abstract. Open USN service is USN service to enable enhanced and flexible service interoperability and provisioning based on the use of standards interfaces. In this paper is to define an open USN service framework, and provide reference architecture of open USN service framework. The use of standard interfaces of open USN service framework will ensure USN service reusability, portability across several USN services, as well as accessibility and interoperability by USN application providers and/or developers. This paper will contribute to the development and activation of new variety USN service by deploying sensor node constructed on variety field or sensor network to share and utilize at different service field.

Keywords: Open USN Service, Semantic, Platform Model, USN, WSN.

1 Introduction

The ubiquitous sensor network (USN) is a well-known keyword in ICT area and many SDOs are developing standards for USN and other similar technologies. However, USN services are not widely spread yet because current USN services require user or application developer to have knowledge of sensors and sensor networks for using USN services or developing USN applications. For example, if a user wants to know the current temperature of Geneva, user should have the information about to which sensors or sensor networks he has to request the sensed data. Also, the user has to know how to interpret sensed data he gets. It is the reason that each USN service use proprietary data format for sensor data, as well as proprietary application programming interfaces (APIs) for developing USN applications [1].

* This work was supported by the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy(MKE, Korea). [10038653, Development of Semantic based Open USN Service Platform].
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_68,

Even with USN middleware which can provide commonly required functionalities by many types of USN applications and services, user has to search, collect, analyse and process the sensed data by himself. Also, many sensed data cannot be shared or reused by users or systems. These are the limits of current USN services and by giving ability of catching the meaning of sensed data and ability of analysing the sensed data these limits can be overcame.

The support of an open USN service framework aims to provide flexible and efficient capabilities base on the use of standard interface to USN applications and services, thereby enabling USN applications to take full advantage of the USN capabilities such as users or application developers do not need any knowledge on sensors or sensor networks to use USN services or develop USN applications. The main purpose of open USN service framework are to provide easy access to the global USN resource and sensed data, installation and connection of low-cost sensors, development and distribution of various applications, and use of the USN resource and sensed data [1].

2 Open USN Service Framework

2.1 Overview of Open USN

Unlike current individual USN service, open USN service opens sensing information, application program interface (API), and USN resources to provide the environment where the user or the service provider can easily use the USN service and develop new services.

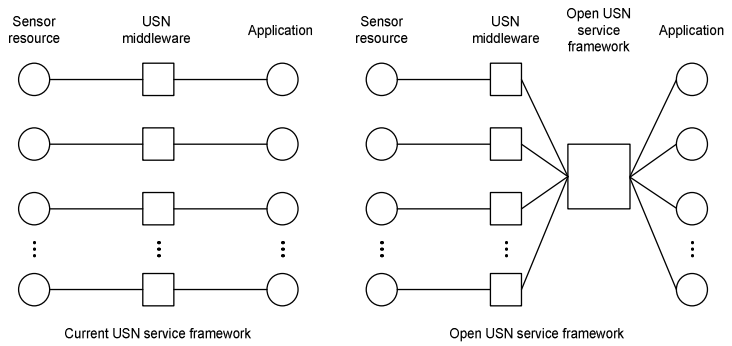


Fig. 1. Current USN service framework and open USN service framework [1]

2.2 Opening Sensing Information

Opening the sensing information collected from USN resources is to share such information for other services. To share sensing information, the sensing information needs to be expressed and defined using semantic technology and so forth. Defined sensing information must be managed by open USN service platform and can be used

to provide additional information in connection with external Linked Open Data (LOD).

2.3 Opening API

The opening of API means providing a single interface to users through standardized API. It avoid APIs that require complicated procedures or the acquisition of new knowledge/technology and avoid direct exposure to the processing method for, expression of, and query for semantic data that users can hardly access, but instead provide the function to intermediate it within open USN service platform in order to allow developers and users can easily access the information or service provided by open USN service using open API.

2.4 Opening USN Resources

Opening USN resources allows users to use USN resources through a variety of service platforms. Generally, the direct user access to USN resources would be blocked due to security reasons, however, sharing through secure service platform would be allows.

3 Proposed Open USN Service Scenario

3.1 Scenario Using Multiple USN Resources

The scenario using multiple USN resources is the service scenario supported by the opening of USN resources. A variety of USN resources can be shared according to the standardized interface between service platform and USN resources, and the composite information can be provided to a specific service by obtaining required information from a variety of USN resources as in (Figure 2). The scenario using multiple USN resource in (Figure 2) is as follows.

- Step 1: The user requests the service platform for a specific service.
- Step 2: The service platform which received user request checks USN resources required to respond to service request and collects sensing information from a variety of USN resources shared through standardized interface
- Step 3: The service platform processes the sensing information collected and response to user.

The user's service request in this scenario can be analyzed and processed by the service platform. Also, the USN resource shared temporarily to process a specific request can be managed in a group (USN resource community) in order to cope with other similar service requests.

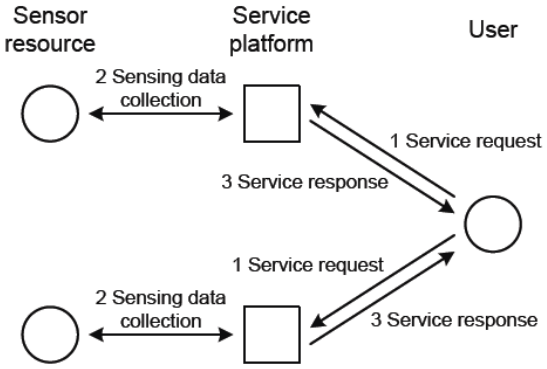


Fig. 2. Scenario using multiple USN resources [2]

3.2 Scenario Using Multiple Service Platforms

The scenario using multiple service platforms is the service scenario supported by open API. The user can use a variety of service platforms through the same API based on the standardized API between service platform and user. In this scenario, the third service provider can easily create and provide a new service by integrating a variety of existing service platforms and USN resources. The scenario using multiple service platforms in (Figure 3) is as follows.

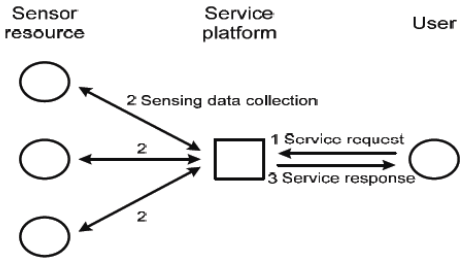


Fig. 3. Scenario using multiple USN resources [2]

- Step 1: The user requests a number of service platforms for service through the standardized API.
- Step 2: After receiving user request, each service platform checks USN resources required to respond to service request and collects sensing information from USN resources through the standardized interface.
- Step 3: Each service platform processes the sensing information collected and response to user.

In this scenario, the user can become the third service provider. In other words, it enables the provision of a new service using a variety of shared service platforms and USN resource.

3.3 USN Resource Registration Scenario

USN resource registration scenario is the scenario to easily register new USN resource to the service platform and provide services. The USN resource registration scenario in (Figure 4) is as follows.

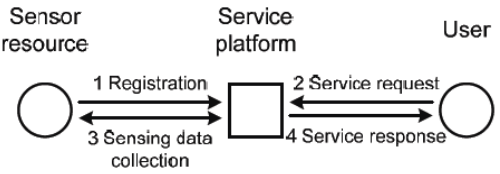


Fig. 4. Scenario using multiple USN resources [2]

- Step 1: When new USN resource is installed, the USN resource is registered to the service platform according to the standardized procedure.
- Step 2: The user requests the service platform for specific service.
- Step 3: The service platform which received user request checks USN resources required to respond to service request and collects sensing information from shared USN resources through standardized interface.
- Step 4: The service platform processes the sensing information collected and responds to user.

This scenario provides the function that allows anyone to install and share USN resources. This allows users to use the individually owned USN resources as though they are social networking services. In this scenario, the plug and play (PnP) based USN resource management function can be used to eliminate complexity in registering and using USN, the biggest constraint in activating USN service.

4 Conclusion

Unlike current individual USN service, open USN service opens sensing information, application program interface (API), and USN resources to provide the environment where the user or the service provider can easily use the USN service and develop new services.

In this paper is to define an open USN service framework, and provide reference architecture of open USN service framework. The use of standard interfaces of open USN service framework will ensure USN service reusability, portability across several USN services, as well as accessibility and interoperability by USN application providers and/or developers. This paper will contribute to the development and activation of new variety USN service by deploying sensor node constructed on variety field or sensor network to share and utilize at different service field.

References

1. Lee, J.S., Kim, S.J.: F.OpenUSN "Requirements and reference architecture for open USN service framework" (New): Initial draft from SG16 meeting (Geneva, 20 November – 2 December 2011). International Telecommunication Union, Study Group 16, TD 697(WP 2/16), Question(s).25/16 (2011)
2. Lee, J., Kim, M., Park, D.: Requirements and Reference Architecture for Open USN Service Framework. Telecommunications Technology Association(TTA) Standard, TTAK.KO-06.0282 (2012)
3. Park, D.-H., Kim, S.-J., Bang, H.-C.: Semantic Open USN Service Platform Architecture. In: ICTC 2012, pp. 12–15 (2012)
4. Kim, M., Lee, J.W., Lee, Y.J., Ryou, J.-C.: COSMOS: A Middleware for Integrated Data Processing overHeterogeneous Sensor Networks. ETRI Journal 30(5), 696–706 (2008)
5. Park, N., Kwak, J., Kim, S., Won, D.H., Kim, H.W.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
6. Park, N.: Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
7. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)
8. Park, N., Kim, Y.: Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS, vol. 6422, pp. 193–202. Springer, Heidelberg (2010)
9. Park, N.: Customized Healthcare Infrastructure Using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
10. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
11. Park, N.: The Implementation of Open Embedded S/W Platform for Secure Mobile RFID Reader. The Journal of Korea Information and Communications Society 35(5), 785–793 (2010)
12. Park, N., Cho, S., Kim, B.-D., Lee, B., Won, D.: Security Enhancement of User Authentication Scheme Using IVEF in Vessel Traffic Service System. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Application. LNEE, vol. 203, pp. 699–705. Springer, Heidelberg (2012)
13. Park, N., Cho, S., Kim, B., et al.: Security Enhancement of User Authentication Scheme Using IVEF in Vessel Traffic Service System. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Application. LNEE, vol. 203, pp. 699–705. Springer, Heidelberg (2012)
14. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
15. Park, N., Ko, Y.: Computer Education's Teaching-Learning Methods Using Educational Programming Language Based on STEAM Education. In: Park, J.J., Zomaya, A., Yeo, S.-S., Sahni, S., et al. (eds.) NPC 2012. LNCS, vol. 7513, pp. 320–327. Springer, Heidelberg (2012)
16. Kim, K., Kim, B.-D., Lee, B., Park, N.: Design and Implementation of IVEF Protocol Using Wireless Communication on Android Mobile Platform. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech/CA/CES3 2012. CCIS, vol. 339, pp. 94–100. Springer, Heidelberg (2012)

Use-Cases and Service Modeling Analysis of Open Ubiquitous Sensor Network Platform in Semantic Environment*

Taegyeong Kang[1], Namje Park[1,**], Hyungkyu Lee[2], and Hyo-Chan Bang[2]

[1] Major in Elementary Computer Education, Department of Primary Education, Graduate School of Education, Jeju National University, 61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Korea
{ktg,namjepark}@jejunu.ac.kr
[2] Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, Korea
{leehk,bangs}@etri.re.kr

Abstract. The ubiquitous sensor network is a well-known keyword in information and communication technology area and many standards development organizations are developing standards for USN and other similar technologies. However, ubiquitous sensor network services are not widely spread yet because current ubiquitous sensor network services require user or application developer to have knowledge of sensors and sensor networks for using ubiquitous sensor network services or developing Ubiquitous Sensor Network applications. In this paper describes Use-cases and Service Modeling Analysis specific to the support of open ubiquitous sensor network service framework.

Keywords: Open USN Service, Semantic, Platform Model, USN, WSN

1 Introduction

In current individual USN service architecture, specific USN resources are dependent on specific service platform and the user is also dependent on the API the specific service platform provides. In individual USN service architecture, there are many limits in using already developed USN resources as the public infrastructure not to mention the inability to use such resources in other services.

For this, it is necessary to develop an open USN service architecture to use the USN resources developed in the past and to be developed in the future as the public infrastructure. The open USN service architecture enables the sharing of USN resources and service platform through the standardized interface between USN

* This work was supported by the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy(MKE, Korea). [10038653, Development of Semantic based Open USN Service Platform].
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_69,

resources and service platforms and the standardized API between service platform and user. Also, it combines shared USN resources to provide multiple services. The support of an open USN service framework aims to provide flexible and efficient capabilities base on the use of standard interface to USN applications and services, thereby enabling USN applications to take full advantage of the USN capabilities such as users or application developers do not need any knowledge on sensors or sensor networks to use USN services or develop USN applications. The main purpose of open USN service framework are to provide easy access to the global USN resource and sensed data, installation and connection of low-cost sensors, development and distribution of various applications, and use of the USN resource and sensed data [1,2].

In this paper describes Use-cases and Service Modeling Analysis specific to the support of open ubiquitous sensor network service framework.

2 Overview of Open USN Service Framework

Open USN Service is a common sensor network middleware platform over heterogeneous USN (Ubiquitous Sensor Network) for USN service applications [2]. Open USN Service provides various functionalities for sensor network service applications. Main functionalities are sensor network abstraction using sensor network common interface, query optimization, integration of data from various sensors, sensor network monitoring, and intelligent sensor data processing such as event handling and sensor data mining [2,3]. Wired and wireless sensor network connect to Open USN Service middleware platform using a sensor network adaptor. Each sensor network adaptor must implement a sensor network abstraction protocol called the sensor network common interface that is proposed as an interface protocol between a sensor network middleware and sensor network field applications in Open USN Service middleware platform [2,3].

Open USN Service operates like a sensor database that contains various sensing values of multiple sensor networks conceptually. It provides an SQL-like query to gather sensing values in sensor network applications. The query processor and optimizer performs the aggregation and filtering of sensing value stream data, and provides user-friendly SQL-like query to service API. The sensor network directory service provides the static metadata and dynamic metadata of sensor network components – sensor networks, sensor nodes, transducers, and related hardware specifications based on sensor network physical configuration. The sensor network dynamic metadata are the link quality between nodes, parent node's id, node's battery level, node's location and so on. The sensor network monitor monitors variable state or values of sensor networks connected to middleware and it updates these dynamic metadata using sensor network directory service's API [2,3].

3 Proposed Open USN Service Framework Model

Open USN service shall be able to provide the following services.

- The service that allows the user to easily access the sensing information without noticing USN technology
- The service that allows the user to easily purchase and install the sensor and connect to the network
- The service that allows ordinary developers to develop a variety of services using the sensing information using open API
- The service that allows anyone to easily use the semantic processed high value sensing information

The open USN service must satisfy the following requirements in order to provide said services

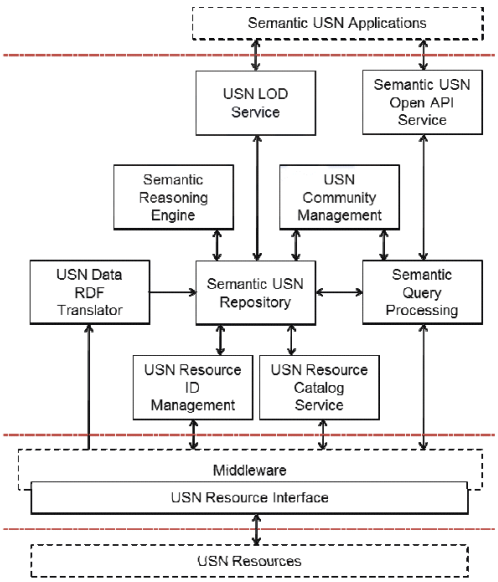


Fig. 1. Reference architecture of open USN service framework [2]

Interface FE providing open USN service framework interfaces to 3rd party USN application and service providers for open USN service. LOD linking FE providing the capability importing external LOD data to open USN service framework. Semantic inference FE providing logical group management for satisfying applications' service requests. Resource group management FE providing management capability on USN resources to grouping or ungrouping for the general purpose. RDF translation FE providing translation capability from raw sensed data collected from USN middleware into semantic USN data format in a form of RDF. Semantic USN repository FE providing general purpose repository for Semantic USN data. Semantic query FE providing query capability (e.g., SPARQL – Query Language for RDF) for Semantic USN repository and USN middleware. Resource ID management FE providing management capability for Resource ID. Resource catalog

FE providing information regarding USN middleware and USN resources. Adaptation FE providing capability to reuse other USN middleware [1].

4 User-Case for Open USN Service Platform

4.1 Use Case for Already Installed USN Resources: Fish Farm Monitoring Service

Figure 2 shows a use case for a fish farm monitoring service when it is assumed that USN resources for this service are previously installed and connected to open USN service framework. The purpose of this service is to provide fish wholesalers with a kind of solution based on open USN service framework to figure out the followings: how to find and monitor tuna farming information in real-time and how to buy tuna raised in the cleanest environment [1].

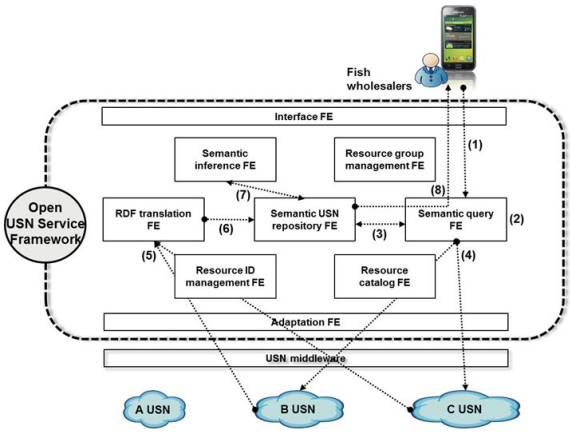


Fig. 2. Use case for already installed USN resources [1]

The steps to accomplish this use case are: 1) Query about a tuna farm monitoring service through the Interface FE by the Application. 2) Interpret query meaning by the Semantic query FE. 3) Discover requested USN resources from the Semantic USN repository FE (assume that B and C USN resources are discovered). 4) Request sensed data to the USN middleware via the Adaptation FE by the Semantic query FE. 5) Collect sensed data from the USN middleware (in this case, from B and C USN resources) via the Adaptation FE. 6) Store sensed data, which are translated into RDF format by the RDF translation FE, on the Semantic USN repository FE. 7) Create inference data by the Semantic inference FE and store these data on the Semantic USN repository FE. 8) Provide contents including sensed and inference data by the Semantic query FE from the Semantic USN repository FE[1].

4.2 Use Case for Dynamically Installed USN Resources: Pet Management

Figure 3 shows a use case for a pet management service when it is assumed that users start to install USN resources for this service.

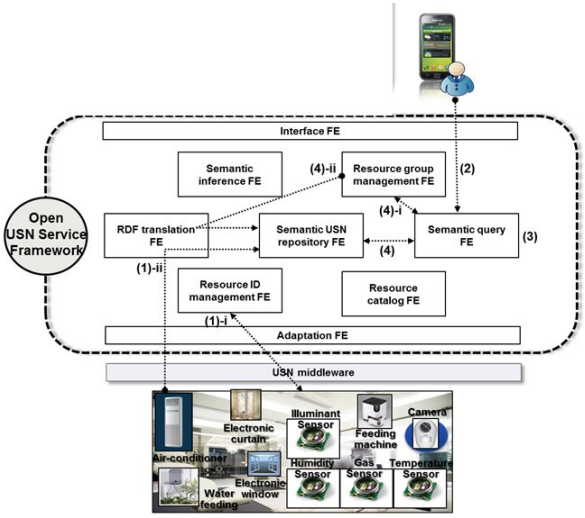


Fig. 3. Use case for dynamically installed USN resources [1]

The purpose of this service is to provide pet owners with a kind of solution based on open USN service framework to solve how to take care of their pet while they are away on a trip for a while.

The steps to accomplish this use case are: 1) Install (or register) USN resources such as sensors (e.g., illuminant, humidity, gas, temperature sensor, camera, etc.) and actuators (e.g., electronic curtain/window, water/feeding machine, air conditioner, etc.) to the open USN service framework by the USN middleware via the Adaptation FE as follows. 2) Assign resource IDs of USN resources by the Resource ID management FE. 3) Store metadata of USN resources collected through USN middleware, which are translated into RDF format by the RDF translation FE, on the Semantic USN repository FE. 4) Query about a pet management service through the Interface FE by the Application. 5) Interpret query meaning by the Semantic query FE. 6) Discover requested USN resources from the Semantic USN repository FE (assume that these USN resources cannot be found). 7) Create logical groups of USN resources by the Resource group management FE. 8) Store these data, which are translated into RDF format by the RDF translation FE, on the Semantic USN repository FE. 9) Remaining steps are same with from (4) to (8) step in the fish farm monitoring service.

5 Conclusion

USN (Ubiquitous sensor network) is one of rising technologies in IT convergence domain. USN is a core technology in environmental monitoring such as air/water quality, radiation and traffic noise. Useful information gathered from sensor network is become to information silo because of USNs are operating tightly coupled with sensor application until now. Sensor data representation formats in these sensor networks are various, and cannot understand the meaning of sensing value in other applications. Therefore, sensing values and information must be shared and provided additional information for other applications [2,3].

In this paper is to define an open USN service framework, and provide reference architecture of open USN service framework. The use of standard interfaces of open USN service framework will ensure USN service reusability, portability across several USN services, as well as accessibility and interoperability by USN application providers and/or developers. The proposed semantic open USN service platform model is supporting semantic expression and interoperability of sensing value and related information, and we can share semantic sensor information in other sensor applications. This paper will contribute to the development and activation of new variety USN service by deploying sensor node constructed on variety field or sensor network to share and utilize at different service field.

References

1. Lee, J.S., Kim, S.J.: F.OpenUSN "Requirements and reference architecture for open USN service framework" (New): Initial draft from SG16 meeting, Geneva, November 20-December 2. International Telecommunication Union, Study Group 16, TD 697(WP 2/16), Question(s).25/16 (2011)
2. Lee, J., Kim, M., Park, D.: Requirements and Reference Architecture for Open USN Service Framework. Telecommunications Technology Association(TTA) Standard, TTAK.KO-06.0282 (2012)
3. Park, D.-H., Kim, S.-J., Bang, H.-C.: Semantic Open USN Service Platform Architecture. In: ICTC 2012, pp. 12–15 (2012)
4. Kim, M., Lee, J.W., Lee, Y.J., Ryou, J.-C.: COSMOS: A Middleware for Integrated Data Processing overHeterogeneous Sensor Networks. ETRI Journal 30(5), 696–706 (2008)
5. Park, N., Kwak, J., Kim, S., Won, D.H., Kim, H.W.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
6. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
7. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)
8. Park, N.: Customized Healthcare Infrastructure Using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
9. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)

A Neural Network Based Simple Weak Learner for Improving Generalization Ability for AdaBoost

Jongjin Won[1,2] and Moonhyun Kim[2,*]

[1] Electronics and Telecommunications Research Institute, Korea
[2] Department of Computer Science Engineering, Sungkyunkwan University, Korea
wonjj@ensec.re.kr, mhkim@ece.skku.ac.kr

Abstract. The performance of ensemble, including AdaBoost, is determined by accuracy and generalization ability. However, the currently available AdaBoost's weak learners mostly show high accuracy but rather low generalization ability. In this paper, we introduce three requirements that weak learners must satisfy in order to improve generalization ability of AdaBoost. Then, we propose w-delta learning rule based neural network(NN) as a weak learner that satisfies those requirements. Through experiments, we show that the proposed method improves generalization ability while maintaining the high accuracy.

Keywords: AdaBoost, Generalization ability, W-delta learning rule.

1 Introduction

In machine learning, supervised learning serves as a way to find a hypothesis that can resolve a given problem. However, it is extremely difficult to find a hypothesis with excellent generalization ability. On the other hand, ensemble utilizes multiple hypotheses in order to generate a better hypothesis. That is, ensemble combines a number of weak classifiers to produce a strong classifier.

AdaBoost [1], which is the most commonly used ensemble, generates a set of weak classifiers called weak hypothesis. Each of these weak classifiers is generated by a weak learner. The decision tree and kernel based NN are mostly used weak learner for Adaboost. However, decision tree is facing with the difficulty of determining an appropriate number of trees and size of each tree that can reduce generalization error. And kernel method is facing with the difficulty of determining an appropriate number of centers and width of kernel that can maximize the classification accuracy. The root cause of these difficulties is that the requirements for weak learners to improve both classification accuracy and generalization ability are not aptly suggested nor sufficiently satisfied. Maintaining a good trade-off relationship between accuracy and generalization ability is a challenge not only for boosting but also for machine learning [2]. When these two elements are balanced well, AdaBoost can finally perform at its full potential. However, the currently available AdaBoost's weak learners are not providing adequate solutions to overcome this challenge.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_70,

In this paper, we analyze requirements for weak learners to improve the generalization ability of AdaBoost and propose a w-delta learning rule based single-layer NN as a weak learner for AdaBoost.

2 Background

2.1 AdaBoost

AdaBoost is a meta learning algorithm that provides a high performance by boosting weak classifiers. AdaBoost's weak learners generate weak classifiers based on weight distribution which shows training samples' level of importance. As shown in Table 1, at each round t weak learner trains weak classifier h_t that minimizes misclassification errors for the given training samples and their weight distribution D_t. This follows the calculation of α_t that signifies the importance of the chosen h_t. As noted in Table 1, we can see that α_t is inversely proportional to h_t's misclassification error ε_t. $D_t(i)$ is decreased if h_t correctly classifies the corresponding training samples and increased if it misclassifies. That is, it assigns a high distribution for the hard examples that remained at round t so that it can successfully select h_{t+1} that allows it to correctly classify them in the subsequent round $t+1$. In the end, the final classifier H is a weighted majority vote of T weak classifiers where α_t is the weight assigned to h_t.

Table 1. AdaBoost algorithm

Given : Set of labeled training samples $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m), where\ \mathbf{x}_i \in \mathbf{X}, y_i \in \{-1, +1\}$

Initialize weight distribution : $D_1(i) = \frac{1}{m}, i = 1, \cdots, m$

For $t = 1, \cdots, T$

(1) Use weak learner to train weak classifier on training samples & distribution D_t

(2) Get weak classifier h_t:$\mathbf{X} \to \{-1, +1\}$ with error $\varepsilon_t = \sum_{i=1}^{m} D_t(i)[y_i \neq h_t(\mathbf{x}_i)]$

(3) If $\varepsilon_t \geq \frac{1}{2}$, then stop, else Set $\alpha_t = \frac{1}{2} ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

(4) Update $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} exp^{-\alpha_t} & if\ h_t(\mathbf{x}_i) = y_i \\ exp^{\alpha_t} & if\ h_t(\mathbf{x}_i) \neq y_i \end{cases}$

Output : the final classifier

$$H(\mathbf{x}) = sign\left(\sum_{i=1}^{T} \alpha_t h_t(\mathbf{x})\right)$$

2.2 Delta Learning Rule

ADALINE is a well-known single-layer NN. The delta learning rule which is described in Eq. (1), applied in ADALINE is a data adaptive technique for deriving a MSE solution which bases on an iterative gradient descent algorithm.

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \eta(y_i - n_i)\mathbf{z}_i \tag{1}$$

3 Requirements for Weak Learner

As an essential requirement for ensemble to ensure increased accuracy of a classifier, Hansen and Salamon have suggested that the classifier must be slightly more accurate than random guesses and must be diverse [3]. What it means by classifier being diverse is that the misclassifying patterns should be different from each other. Moreover, for weak learners to be utilized by AdaBoost more effectively, weak learners must be able to generate weak classifiers based on training samples' distribution.

Accuracy. Wickramaratna et al. suggested that AdaBoost can provide its maximum performance when weak learners are moderately weak [4]. AdaBoost is designed to improve overall performance by using committee of weak classifiers that learn from different characteristics of training samples in training data space. However, because of the fact that AdaBoost gives more emphasis on hard samples and outliers, it tends to over-fit training examples. If hard examples and outliers - which can be a small portion of the whole - are overly emphasized, the rest of rather important samples' features can be overlooked and result in decreased performance of the final classifier.

Diversity. Learning algorithm is said to be unstable when a small change in training samples brings a relatively much impact on classification accuracy [5]. C4.5 [6] which is used to generate decision tree is a prime example. On the other hand, learning algorithm is said to be stable when a small change in training samples hardly impacts on its classification accuracy (e.g., SVM [7]). Ensemble can be used to increase the performance of classifiers generated from unstable algorithm. However, for classifiers generated from stable algorithm [8], some studies have shown that ensemble can also be used to increase their accuracy but the performance improvement only applies to the limited type of data, compared to the unstable algorithms.

The reason why unstable algorithm, which generally shows poor performance, has a room to improve the classification ability through boosting is that it has ensured a wide variety of weak classifiers while stable algorithm is restricted from increasing the classification ability due to lack of diversity. When classifiers that comprise ensemble are diverse, even if a classifier misclassifies a specific sample, since other classifiers can correctly classify the sample, in the end, it can be correctly classified through ensemble.

Use of Samples' Distribution. As previously mentioned, weak learner must take the samples' distribution into account. This can be done in two ways - reweighting and resampling [1]. Some studies have shown that resampling produces better performance than reweighting. However, we believe that resampling is less efficient than reweighting in a way that it must resample according to samples' distribution after each iteration. Although it is possible that weak learner which cannot handle samples' distribution can be applied to AdaBoost, it is suggested to use weak learner which can handle samples' distribution as it provides better computational efficiency.

4 Proposed Weak Learner

4.1 W-delta Learning Rule

As shown in Eq. (1), the learning rate(η) of delta learning rule plays an important role of determining how much of given samples to apply to learning. If η is too large, delta learning rule can experience divergent oscillations and increase errors. If η is too small, the learning itself can take up very long computation times [9].

Table 2. W-delta learning algorithm

Given : Set of labeled training samples with distribution $(\mathbf{x}_1, y_1, D_1), \cdots, (\mathbf{x}_m, y_m, D_m)$

(1) Calculate largest eigenvalue λ_{max} of training samples

(2) Transform D_i to optimal learning rate $D_i' = f(D_i)$

$$where\ f: [D_{min}, \cdots, D_{max}] \rightarrow \left[0, \cdots, \frac{2}{\lambda_{max}}\right]$$

(3) Train for $T^{w-delta}$ times for all training samples

$$\mathbf{w}_{i+1} = \mathbf{w}_i + D_i'(y_i - n_i)\mathbf{z}_i$$

Output : trained neuron's weight $\mathbf{w}$

Since delta learning rule is based on the steepest decent algorithm, the maximum stable learning rate is $\eta < \frac{2}{\lambda_{max}}$ and it uses a fixed η for the duration of training process [10]. Here, λ_{max} is the largest eigenvalue of training samples' covariance matrix.

If samples and samples' distribution are provided, the suggested w-delta learning rule calculates λ_{max} to find a stable learning rate. Then the function f (in Table 2) transforms the samples' distribution so that it falls into the range of stable learning rate. We adopt log-sigmoid function to this transformation in order to make w-delta learning rule produce loosely correlated weak classifiers. Since log-sigmoid function emphasizes on high distribution more than low distribution, the w-delta learning rule reflects the samples which previous classifiers misclassified to learning as much as possible and, at the same time, applies the samples which previous weak classifiers correctly classified to learning as little as possible. Therefore, it generates loosely correlated classifiers. The function f is described in Eq. (2), where β is the slope parameter and $\bar{\mathbf{x}}$ is the mean vector of training samples.

$$f(x) = \frac{2}{\lambda_{max}}\left(\frac{1}{1 + e^{-\beta(x-\bar{\mathbf{x}})}}\right) \tag{2}$$

Also, since the transformed samples' distribution D' is always less than maximum stable learning rate, it prevents from causing divergent oscillations. If w-delta learning rule is used as a weak learner, it can generate weak classifiers that can classify samples with high distribution properly.

4.2 Improvement of Generalization Ability

Unlike Table 1, AdaBoost that uses w-delta learning rule does not set the number of weak classifiers(T) in advance. Since w-delta learning rule generates weak classifiers using the actual samples' distribution, as rounds continue to repeat, it can generate weak classifiers that classify samples with high distribution well.

The weak classifier that is generated at the point where AdaBoost over-fits samples has a high chance of misclassifying the most of training samples with low distribution. Such weak classifiers will have $\varepsilon_t \geq \frac{1}{2}$ and end algorithm. Hence, due to the fact that w-delta learning rule uses the samples' distribution, AdaBoost can produce classifiers with high generalization ability (i.e., AdaBoost does not over-fit samples).

5 Experiment and Result

In this section, we compare the suggested w-delta learning rule based AdaBoost($AdaBoost_{w-delta}$) with decision tree based AdaBoost($AdaBoost_{DT}$). Ten benchmark data sets from UCI Repository and STATLOG are used to evaluate the generalization ability and accuracy of the proposed algorithm.

Table 3. The experimental result of $AdaBoost_{DT}$ and $AdaBoost_{w\text{-}delta}$

Data set	% Correct (# Weak classifiers)			
	$AdaBoost_{DT}$			$AdaBoost_{w\text{-}delta}$
	Splits # = 2	Splits # = 3	Splits # = 4	
australian	**85.5(50)**	83.7(100)	83.1(5)	**85.2(5)**
b. cancer	94.1(20)	95.3(100)	**95.9(50)**	**98.8(5)**
diabetes	**72.1(5)**	71.1(100)	68.9(20)	**79.4(4)**
german	**75.6(20)**	74.8(5)	75.2(50)	**74.6(13)**
heart	**76.1(10)**	**76.1(20)**	**76.1(50)**	**85.2(6)**
ionosphere	90.9(100)	**94.9(50)**	91.4(20)	**96.0(8)**
l. disorders	**72.9(50)**	68.2(10)	69.4(100)	**66.3(5)**
spambase	92.4(100)	**93.0(50)**	92.9(100)	**88.8(24)**
spect	69.1(5)	**71.1(5)**	68.0(5)	**79.4(4)**
s. junctions	**92.6(100)**	92.4(50)	91.4(5)	**93.1(8)**

In Table 3, the result of applying $AdaBoost_{DT}$ for the data set show that it is difficult to find an appropriate number of trees and splits that indicates the highest accuracy for each data set. For most of data set, it shows a high accuracy when the number of trees (number of weak classifiers) is at least 10 or more. For $AdaBoost_{w\text{-}delta}$, we did not use a fixed number of weak classifiers due to the characteristic of the suggested method. We can confirm that the generalization ability is improved while keeping the number of weak classifiers below 10 for most data sets and maintaining the similar accuracy compared to the result of $AdaBoost_{DT}$.

If weak classifiers are tightly correlated, their misclassifying patterns are similar to each other. This means that a final classifier will need a number of weak classifiers. It will only need a much smaller number of weak classifiers, if their misclassifying

patterns are uncorrelated. AdaBoost has a tendency of over-fitting data. This causes generating excessive number of weak classifiers in order to properly classify training samples. Conversely, because the method proposed in this paper uses NN to train samples that base on distribution, it was possible to produce the similar accuracy to the method of using decision tree even with a small number of weak classifiers. Diversity of decision tree and NN is similar because they are unstable algorithms. But, in our proposed method, we use the nature of single-layer NN where its decision boundary is directly determined by training sample. Therefore, we can confirm that the suggested method improves the generalization ability of AdaBoost using a small number of classifiers as shown through the experiment.

6 Conclusions

In this paper, we identified three requirements that weak learners must satisfy to improve the performance of AdaBoost. Then we designed w-delta learning rule that overcomes the problem of existing weak learners and suggested w-delta leaning rule based single-layer NN as AdaBoost's weak learner.

Among the requirements mentioned above, the diversity is the most important factor for weak learner. So, we adopt single-layer NN to avoid over-fitting. The proposed method has more diversity than other studies because w-delta learning is able to generate loosely correlated weak classifiers through uncorrelated training samples and transformed distribution. Conclusively we described how the proposed method satisfied the identified requirements for weak learner and, through experiment, we confirmed that it maintained the same high accuracy as the existing methods and yet improved generalization ability significantly.

References

1. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139 (1997)
2. Levin, E., Tishby, N., Solla, S.A.: A statistical approach to learning and generalization in layered neural networks. Proc. of IEEE 78(10), 1568–1574 (1990)
3. Hansen, L., Salamon, P.: Neural network ensembles. IEEE Trans. PAMI 12(10), 993–1001 (1990)
4. Wickramaratna, J., Holden, S., Buxton, B.: Performance degradation in boosting. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 11–21. Springer, Heidelberg (2001)
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd ed. Wiley (2000)
6. Quinlan, J.R.: C4.5 programs for machine learning. Morgan Kaufmann Publishers (1993)
7. Vapnik, V.: Statistical Learning Theory. John Wiley (1998)
8. Evgeniou, T., Perez-Breva, L., Pontil, M., Poggio, T.: Bound on the generalization performance of kernel machine ensembles. In: Proc. of ICML, pp. 271–278 (2000)
9. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press (1995)
10. Hagan, M.T., Demuth, H.B., Beale, M.: Neural Network Design. PWS (1996)

Serial Dictatorial Rule-Based Games for Camera Selection

Gowun Jeong[1], Yong-Ho Seo[2,*], Sang-Soo Yeo[3], and Hyun S. Yang[1]

[1] AI and Media Lab., Dept. Computer Science, KAIST, Daejeon, Republic of Korea
{gowun,hsyang}@kaist.ac.kr
[2] Dept. Intelligent Robot Engineering, Mokwon University, Daejeon, Republic of Korea
yhseo@mokwon.ac.kr
[3] Division of Computer Engineering, Mokwon University, Daejeon, Republic of Korea
ssyeo@ieee.org

Abstract. A wireless, battery-powered, stationary camera sensor network optimizes trade-off between extending its lifetime and enhancing its sensing accuracy by activating only a desirable camera subset for given targets in a timely fashion. This paper models this selection problem in a cooperative bargaining game based on the serial dictatorial rule, where cooperative sensors sequentially decide their mode between "sleep" and "active" in descending order of their bargaining power. Simulated resource overheads and the concerned performances, network lifetime and sensing accuracy are given as well.

Keywords: sensor scheduling, coverage, cooperative bargaining game, serial dictatorial rule.

1 Introduction

Since the demand to cost-efficiently extract meaningful information about multi-targets moving inside a wide area has increased in surveillance areas, networks of wireless, battery-powered, field-of-view (FoV) overlapping, stationary camera sensors have been widely employed due to such sensors' cheapness and handy installation [1]. The sensors more seriously suffer from their limited capabilities on bandwidth, lifespan, computation, and storage than other wireless sensors do because they deal with not only 2D points of the locations of given targets but high-dimension data sets of their richer information from images [2]. Thus, it is required to efficiently activate only selected cameras which collectively optimize the coverage of given targets in a timely fashion, called *Camera Selection* (CS).

To solve this CS problem, this paper aims at that every camera sensor can autonomously select its mode between "sleep" and "active" only with local knowledge to optimize the collective coverage of given targets for advanced target analysis, beyond just tracking. As in [3, 4], each of multiple parties, cameras in the paper, should cooperatively bargain for an optimal collective coverage. Our CS utilizes the serial

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_71,

dictatorial rule, by which more preferred cameras are prioritized to select their mode in earlier steps, for efficient computation. The remainder discusses not only our proposed serial dictatorial rule-based cooperative game for CS in detail but its required overheads and resulted network performances in multi-camera multi-target simulations.

2 Game Modeling for CS

Consider c cameras, indexed i, statically deployed and t targets, indexed j, randomly moving inside a geographical area. We also assume that any two cameras directly communicate with each other if their FoVs are overlapping. We call these cameras *neighbors*. The locations of cameras are initially calibrated, whereas those of targets are speculated by any object localization algorithm of [5] as needed and then are known to their associated cameras for successive target location estimation. Given estimated locations of targets, every camera i expects its set of observable targets in its FoV, T_i.

Saying that m_i represents the mode of camera i between 0 for sleep and 1 for active, we formulate the target utility of target j by i given m_i, the camera utility of i given m_i, and the global target utility of j given $\{m_i\}_j$, the mode set of cameras whose FoV contains j as in (1), (3), and (4), respectively, as follows.

$$tu_j(m_i) = m_i \min_{j' \in T_i \setminus j} dist(\theta_{jj'}) \tag{1}$$

$$dist(\theta_{jj'}) = \begin{cases} 1 & if\ \theta_{jj'} \geq \frac{90}{A} \\ \sin(A\theta_{jj'}) & if\ \theta_{jj'} < \frac{90}{A} \end{cases} \tag{2}$$

$$cu_i(m_i) = \sum_{j \in T_i} tu_j(m_i) \tag{3}$$

$$U_j(\{m_i\}_j) = \sum_{\{m_i\}_j} tu_j(m_i) \quad (\leq 1) \tag{4}$$

Unlike scalar sensors, camera sensors have the following particular characteristics.

i. *Due to the 3D-to-2D projection of photographing, the occlusion amongst multiple targets in one's FoV obstructs extracting the targets' information* [2, 3].
ii. *The interaction analysis amongst multiple targets provides more meaningful information about the targets beyond their locations* [6].

Equations (1) and (3) respectively represent i. and ii. More specifically, j's target utility of (1) stands for the least likelihood that j is observed without any occlusion in i's FoV according to i's mode. Since the greater $\theta_{jj'}$, the more alleviated the occlusion between j and j', we define the distinction probability between images

simultaneously taken by j and j', as (2) for $\theta_{jj'}$, the included angle between the i-to-j vector and the i-to-j' vector, and A, the scaling factor by relying on a sine curve. To evaluate i's contribution for the interaction analysis of T_i according to m_i, i's camera utility is given as the sum of every j's target utility over $j \in T_i$ as in (3). Given all the modes of every camera whose FoV contains j, the likelihood that j is observed by the cameras in total is employed as j's global target utility by (4), whose upper bound is set to 1. Then, all the cameras cooperatively tend to maximize their payoff, the global utility defined as $U_g(m_i) = \sum_{j=1}^{t} U_j(\{m_i\}_j)$ for appropriate $\{m_i\}$ for every camera i.

3 Serial Dictatorial Rule-Based Bargaining Solution

According to [7], the serial dictatorial rule is a sequence of dictatorial rules conducted by individual players whose exercising order is statically arranged by their bargaining powers. By evaluating camera utilities given as in (3), we consider that a camera is more bargaining-powerful than another if it observes more targets less occluded in the corresponding safe region. Given every mode of its more bargaining-powerful cameras determined, when a camera faces its order to choose a mode, the one maximizing its payoff, that is U_g, is selected by the dictatorial rule under the assumption that the others which have not decided their modes sleep. This bargaining process is serially carried out until every camera decides its mode with communicating with neighboring cameras as follows.

1. Order cameras by their camera utilities.
   Given estimated locations of T_i, every camera i computes the target utilities of every target in T_i and its own camera utility assuming $m_i = 1$ and share them with its neighboring cameras. Then, i finds its position in the dictatorial ordering list of it and its neighbors in descending order of their camera utilities, while initializing the mode set for the list, $M_i = \{0\}$.
2. Select the current mode.
   Prior to mode selection, every i waits for all the modes of the more bargaining-powerful neighbors to be announced while updating M_i if it is not the first of the list. Otherwise, it instantly takes its mode by (5) for $M_{i,j}^{m_i}$ is the target j's associated subset of M_i where only the mode of i is replaced by m_i.

$$m_i = \begin{cases} 1 & if \quad \sum_{j \in T_i} \left(U_j\left(M_{i,j}^{1}\right) - U_j\left(M_{i,j}^{0}\right) \right) > 0 \\ 0 & otherwise \end{cases} \tag{5}$$

   Only if i can improve the total of the global target utilities for every target of T_i by its contribution, it decides to be active. In other words, *a camera covers its FoV only if at least one target in its FoV is not sufficiently observed for the moment*, which means that it needs not to be responsible for the targets already well-covered by other cameras. After this mode decision, i announces its selected mode followed by that i and every neighbor, termed i', updates $M_{i'}$ with i's new mode and drops every j such that $U_j\left(M_{i',j}\right) = 1$ obtained by the new $M_{i'}$ from $T_{i'}$ for rapid computation.

This local reasoning with limited knowledge soundly and completely extends maximizing U_g for the following Lemma.

Lemma 1. For every camera i with its mode m_i, it always holds that

$$U_g(M^*) - U_g(M^{*,-i}) \geq \sum_{j \in T_i} (U_j\left(M_{i,j}^{m_i}\right) - U_j\left(M_{i,j}^{m_{-i}}\right))$$

where M^* is the bargained mode set for every camera, m_{-i} is the opposite mode of m_i, $M^{*,-i}$ is the mode set where only m_i is replaced by m_{-i} in M^*, and $M_{i,j}$ is i's assumed mode set of i and i's neighbors associated with target j for the bargaining process.

Proof. The following equation (6) *holds by the global utility definition, and* (7) *is derived since the change of camera i's mode effects only the global target utilities of every* $j \in T_i$. *As* M_j^* *and* $M_j^{*,-i}$ *are respectively restated as* $M_{i,j}^{m_i}$ *and* $M_{i,j}^{m_{-i}}$, *we need to show that* (8) *always holds for every i with* T_i *for our claim in the following two different cases.*

$$U_g(M^*) - U_g(M^{*,-i}) = \sum_{j=1}^{t} (U_j\left(M_j^*\right) - U_j\left(M_j^{*,-i}\right)) \tag{6}$$

$$= \sum_{j \in T_i} (U_j\left(M_j^*\right) - U_j\left(M_j^{*,-i}\right)) \tag{7}$$

$$\sum_{j \in T_i} (U_j\left(M_j^*\right) - U_j\left(M_j^{*,-i}\right)) \geq \sum_{j \in T_i} (U_j\left(M_{i,j}^{m_i}\right) - U_j\left(M_{i,j}^{m_{-i}}\right)) \tag{8}$$

(a) *Case of* $(m_i, m_{-i}) = (1, 0)$
When i decides its mode as active with assuming that every mode for its less bargaining-powerful cameras in its neighbors is sleep, it believes that it can improving the global target utility of any in T_i. *Let us say that* $\{j'\}$ *is the target set each global target utility of which is actually raised by i. The difference by the i's mode change is given as follows.*

$$\sum_{\{j'\}} \sum_{m_{i'} \in M_{j'} \setminus m_i} tu_{j'}(m_{i'}) \tag{9}$$

Since any of less bargaining-powerful neighbors could be active to contribute to improve the concerned global target utilities, $M_{j'}^*$ *is likely to have more or equal active cameras than* $M_{i,j'}$. *Thus,* (8) *is valid. Also,* (9) *is always greater than or equal to* 0 *by the definition of the target utility, which eventually leads to that* $U_g(M^*) \geq U_g(M^{*,-i})$ *is valid.*

(b) *Case of* $(m_i, m_{-i}) = (0, 1)$
When i decides its mode as sleep, it believes that every $j \in T_i$ *is sufficiently covered by more bargaining-powerful neighbors, which derives*

$\sum_{j\in T_i}\left(U_j\left(M_{i,j}^1\right)-U_j\left(M_{i,j}^0\right)\right)=0$ *by (8). Similarly, it holds that* $\sum_{j\in T_i}\left(U_j\left(M_{i,j}^*\right)-U_j\left(M_{i,j}^{*,-i}\right)\right)=0$ *regardless of the other modes in* M_j^*. *Thus,* (8) *as well as* $U_g(M^*)=U_g(M^{*,-i})$ *is valid in this case, too.*

Therefore, the selected mode by (5) with limited knowledge for every camera optimizes the global utility by Lemma 1.

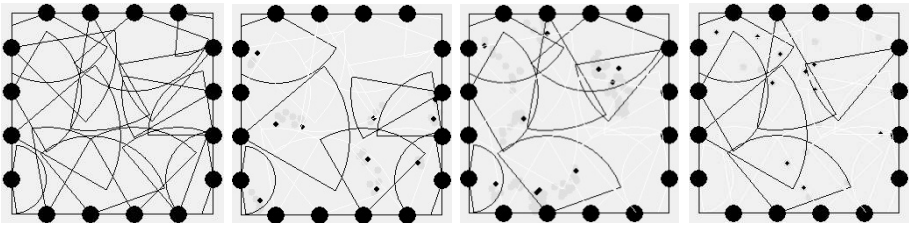
4 Simulations and Analysis

4.1 Simulation Environment and Results

We test our proposed approach in the following simulation environment.

- 16 stationary cameras of $A = 2$ are deployed under the maximum neighboring density $d = 7$ and calibrated in a square-shape area of 230 x 230 cells as in Fig. 1a.
- In a scenario, 10 targets freely move inside the area under speed of 15 cells per time for 1000 times.
- Given target locations at the previous time instance, every camera cooperatively computes the expected current target locations.

Fig. 1 presents three examples of two successes and one failure with the initial camera deployment. Whereas distantly located targets in general are covered as in Fig. 1b, Cameras 3 and 5 collectively monitor their observable targets closely located as in Fig. 1c because they could be occluded in the FoV of Camera 5. Due to wrongly given previous or differently estimated current locations of the targets, cameras may miss targets as Cameras 7 and 9 or be redundantly active as Camera 15 in Fig. 1d.



(a) Cameras with FoVs (b) Distantly located targets (c) Closely located targets (d) Error case

Fig. 1. Camera deployment and three examples of two successes and a single failure with 16 cameras and 10 targets

After 100 random scenario tests, we have obtained the average numbers of active cameras, of missing targets, and of redundantly active cameras as 6.2, 0.32, and 0.85. Defining the network lifetime as the time that has elapsed until targets are covered by

at least one sensor [8], we can expect that a network lives 2.6 (=16/6.2) times longer on average by ours than without any selection technique. Also, 0.32 missing targets and 0.85 redundantly active cameras on average indicate that our proposed approach works well over wrong location estimation, which necessarily occur in any existing localization techniques.

4.2 Complexity Analysis

Since we consider wireless cameras, we cannot help discussing the resource overheads required by our proposed method. By our design, for each time instance, energy consumption happens by the following operations:

1) sensing and image processing; 2) processed data transmitting; 3) estimating target locations at the current time instance; 4) exchanging target locations by broadcasting; 5) computing utilities; 6) exchanging utilities by broadcasting; 7) selecting the mode; and 8) exchanging selected modes by broadcasting.

Operations 1) and 2) are conducted by only active cameras and ours activates 0.39 (=6.2/16) times less cameras as it extends the network lifetime. O(td) communication is dissipated for 4) and 8), and the computation complexity for 3) is known as O(t^2) [9]. Each target utility, each camera utility, and each global target utility respectively consumes O(t), O(t^2), and O(td) computation, which leads to O(t^2+td) computation for 5) and the same amount of communication for 6). Given such utilities, a camera decides its mode by searching in the O(t) space for 7).

The energy consumption of camera sensors is dominated by 1) and 2) due to the enormous size of image data [2, 10]. Thus, ours can be viewed as fairly competitive if targets are not highly crowded in any of FoVs.

5 Conclusion

The paper models a CS in a cooperative bargaining game, where every participating camera serially optimizes the global utility only with local knowledge based on the serial dictatorial rule. The simulated results have presented that ours extends the network lifetime and works well over limitedly accurate target locations as well as that ours is energy-efficient for uncrowded targets.

Acknowledgements. This work was supported by the Basic Science Research Program through a grant from the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0005793, 2011-0014394 and 2011-0013776).

References

1. Akyildiz, I.F., Molodia, T., Chowdhury, K.R.: A survey on wireless multimedia sensor networks. Computer Networks 51, 921–960 (2007)

2. Soro, S., Heinzelman, W.: A survey of visual sensor networks. Advances in Multimedia (2009)
3. Li, Y., Bhanu, B.: Utility-based dynamic camera assignment and hand-off in a video network. In: the ACM/IEEE International Conference on Distributed Smart Cameras, pp. 1–9 (2008)
4. Soto, C., Song, B., Roy-Chowdhury, A.K.: Distributed multi-target tracking in a self-configuring camera network. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009) (2009)
5. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Efficient subwindow search: a branch and bound framework for object localization. IEEE Trans. Pattern Analysis and Machine Intelligence 31(12), 2129–2142 (2009)
6. Candamo, J., Shereve, M., Goldof, D.B., Sapper, D.B., Kasturi, R.: Understanding transit scenes: a survey on human behavior-recognition algorithm. IEEE Trans. Intelligent Transportation Systems 11(1), 206–224 (2010)
7. Kibris, O.: Cooperative game theory approaches to negotiation. In: Handbook of Group Decision and Negotiation. Springer (2010)
8. Pyun, S.Y., Cho, D.H.: Power-saving scheduling for multiple-target coverage in wireless sensor networks. IEEE Trans. Communications Letters 13(2), 130–132 (2009)
9. Triebel, R.: 6. Online Estimation: the Kalman Filter. In: Information Processing in Robotics. ETH (2009)
10. Margi, C.B., Mnduchi, R., Obraczka, K.: Energy consumption tradeoffs in visual sensor networks. In: The 24th Brazilian Symposium on Computer Networks (2006)

Security Analysis on a Group Key Transfer Protocol Based on Secret Sharing*

Mijin Kim[1], Namje Park[2], and Dongho Won[1,**]

[1] College of Information and Communication Engineering, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon, Gyeonggi-do, 440-746, Korea
{mjkim,dhwon}@security.re.kr
[2] Department of Computer Education Teachers College, Jeju National University, Jeju, Korea
namjepark@jejunu.ac.kr

Abstract. Group key exchange protocols are cryptographic algorithms that describe how a group of parties can communicate with their common secret key over insecure public networks. In 2013, Olimid proposed an improved group key transfer protocol based on secret sharing, and claimed that he eliminated the flaws in Sun et al.'s group key transfer protocol. However, our analysis shows that the protocol is still vulnerable to outsider and insider attacks and does not provide known key security. In this paper, we show a detailed analysis of flaws in the protocol.

Keywords: key exchange protocol, group key transfer, secret sharing, attack, confidentiality.

1 Introduction

Secure group communications over public networks require that all group participants have to share a common secret key. This shared secret key, called the session key, is used to expedite authentication, confidentiality, and data integrity services. Group key transfer protocols are designed to achieve the fundamental security goal that no one except the group participants can establish the session key. Over the years, various protocols [1,2,3,4,5,6,7,8,9] have been proposed to achieve the fundamental goal of securely distributing a session key among a group of n participants.

Recently, Sun et al. presented a group key transfer protocol based on secret sharing instead of encryption algorithm [8]. The protocol only needs the server to broadcast $n+1$ messages at once in a round of distribution and all of the legal users only need to store one secret share in all conversations regardless of new addition or someone's

* This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3007) supervised by the NIPA(National IT Industry Promotion Agency).

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_72,

walkout. In addition, a simple computation is enough for each user to obtain the key. However, due to a flaw in Sun et al.'s protocol design, the protocol fails to achieve the fundamental security goal. In 2013, Olimid showed that Sun et al.'s protocol is susceptible to insider attacks and violates known key security and proposed an improved version of the protocol that eliminated the flaws of the original protocol. In this work, we provide a security analysis on the improved group key transfer protocol. Our analysis shows that the protocol still has flaws in the design and can be easily attacked. We present insider attack, outsider attack and failure of known key security on the protocol.

This paper is organized as follows: Section 2 reviews Olimid's group key transfer protocol. Section 3 presents security analysis of the protocol. Finally, Section 4 concludes this work.

2 Olimid's Group Key Transfer Protocol

This section reviews an improved group key transfer protocol [9]. The protocol assumes a trusted key generation center (KGC) who provides key distribution service to its registered users, and consists of two phases: user registration, group key generation and distribution. The protocol adopts the following derivative secret sharing scheme.

Derivative Secret Sharing

Phase 1: Secret sharing

1. KGC splits S into two parts n times: $S = s_1 + s_1' = s_2 + s_2' = \cdots = s_n + s_n'$.
2. KGC sends P_i the share s_i', $i=1,2,\ldots,n$, respectively in a secure channel.

Phase 2: Reconstruction

1. KGC broadcasts the shares $s_i, i = 1,2, \ldots, n$, at once when users want to recover the secret.
2. P_i regains S by computing $S = s_i + s_i'$.

The derivative secret sharing reduces the mutual dependence on others. Detailed steps of these phases are described as follows.

2.1 Olimid's Protocol

Let U be a set of all users who can participate in the protocol. The users in any subset of U may run the protocol to establish common session key.

Phase 1: User registration: Each user is requested to login to KGC for subscribing the group key distribution service. During registration, KGC shares a long-term secret s_i' with each user $U_i \in$U .

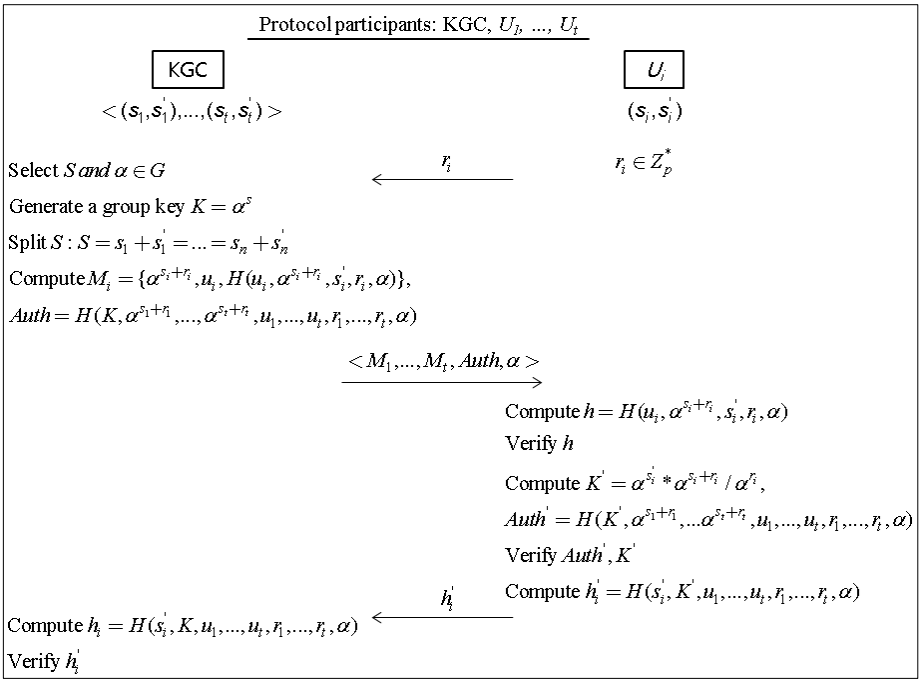


Fig. 1. An execution of Olimid's protocol (described from Step 3)

Phase 2: Group Key Generation and Distribution:

1. The initiator, a designated user of the group, requests for a group key distribution service by sending KGC $\{u_1, u_2, \dots, u_t\}$, which contains the identities of the registered users $U_1, U_2, \dots, U_t$, in current session.
2. KGC broadcasts the list of all participants according to the above received message as a response.
3. Each U_i, $i=1, \dots, t$ sends a random challenge r_i to KGC.
4. KGC randomly selects S and α to generate the group key $K = \alpha^S$ for current service and then invokes derivative secret sharing to split S into two parts t times such that $S = s_1 + s_1^{'} = s_2 + s_2^{'} = \cdots = s_t + s_t^{'}$. KGC then computes: $M_i = \{\alpha^{s_i+r_i}, u_i, H(u_i, \alpha^{s_i+r_i}, s_i^{'}, r_i, \alpha)\}$ and $Auth=H(K, \alpha^{s_1+r_1}, \dots, \alpha^{s_t+r_t}, u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$. At last, KGC broadcasts $\{M_1, \dots, M_t, Auth, \alpha\}$ to the users at once.
5. After receiving $M_i, Auth$, and α, U_i computes $h = H(u_i, \alpha^{s_i+r_i}, s_i^{'}, r_i, \alpha)$, where $\alpha^{s_i+r_i}$ and u_i are from M_i, $s_i^{'}$ is the shared long-term secret stored by U_i, r_i as chosen in step 3. And then U_i checks whether or not h is equal to the corresponding part in M_i. If any of the checks fails, U_i aborts; Otherwise, U_i computes $K^{'} = \alpha^{s_i^{'}} * \alpha^{s_i+r_i}/\alpha^{r_i}$, $Auth^{'} = H(K^{'}, \alpha^{s_1+r_1}, \dots, \alpha^{s_t+r_t}, u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$ and checks whether or not $Auth^{'}$ is equal to $Auth$. If so, then $K^{'}$ is the correct group session key K which is distributed by KGC.

6. Each user U_i returns a value $h_i^{'} = H(s_i^{'}, K^{'}, u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$ to KGC. KGC computes $h_i = H(s_i^{'}, K, u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$ with its own $s_i^{'}$ and K, and checks whether or not $h_i^{'} = h_i$. This review confirms that every user in current session has obtained the correct group key.

3 Security Analysis

In this section, we analyze the security features of the improved group key transfer protocol based on secret sharing described in Section 2. The fundamental security goal of a key exchange protocol is to ensure that no one other than the intended users can compute the session key. But, Olimid's protocol fails to achieve this fundamental security goal. We describe this security vulnerability of the improved group key transfer protocol.

3.1 Outsider Attack

To an outside adversary, his motivation is to obtain the group key or share the group key with group participants. In the following analysis, we can see that his aim is come true.

Method 1:

1. The adversary A can grasp $r_i, M_1, \dots, M_t, Auth,$ and α from the broadcast channel between KGC and authorized users U_i.
2. Since A knows $M_i = \{\alpha^{s_i+r_i}, u_i, H(u_i, \alpha^{s_i+r_i}, s_i^{'}, r_i, \alpha)\}$ $(i = 1,2, \dots, t)$, A is able to obtain $H(u_i, \alpha^{s_i+r_i}, s_i^{'}, r_i, \alpha)$.
3. Using the grasped values $r_i, \alpha^{s_i+r_i}, u_i,$ and α, A is able to obtain $s_i^{'}$ from $H(u_i, \alpha^{s_i+r_i}, s_i^{'}, r_i, \alpha)$ by guessing attack.
4. Thus, A is able to calculate the session key $K = \alpha^{s_i^{'}} * \alpha^{s_i+r_i}/\alpha^{r_i}$.

Method 2:

1. A can grasp $r_i, M_1, \dots, M_2, Auth,$ and α from the broadcast channel between KGC and authorized users U_i.
2. From $M_i = \{\alpha^{s_i+r_i}, u_i, H(u_i, \alpha^{s_i+r_i}, s_i^{'}, r_i, \alpha)\}$ $(i = 1,2, \dots, t)$, A is able to obtain u_i and $\alpha^{s_i+r_i}$.
3. Using the grasped values $Auth$, $\alpha^{s_1+r_1}, \dots, \alpha^{s_t+r_t}, u_1, \dots, u_t, r_1, \dots, r_t,$ and α, A is able to obtain the session key K From $Auth = H(K, \alpha^{s_1+r_1}, \dots, \alpha^{s_t+r_t}, u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$ by launching a guessing attack.

3.2 Insider Attack

Every inside user in Olimid's protocol is expected to reconstruct the group key but know nothing more extra information. However, our analyses show that malicious inside user P_i can forge the return response $H(s_j^{'}, K^{'}, u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$ and impersonate P_j as following.

Method 1:

1. A can grasp $r_1, \dots, r_t, M_1, \dots, M_t, Auth,$ and α from the broadcast channel between KGC and authorized users U_i.
2. Since A knows $M_j = \{\alpha^{s_j+r_j}, u_j, H(u_j, \alpha^{s_j+r_j}, s_j', r_j, \alpha)\}$ $(i = 1,2, \dots, t)$, A is able to obtain $H(u_j, \alpha^{s_j+r_j}, s_j', r_j, \alpha)$.
3. Then, A knows $r_j, u_j, \alpha^{s_j+r_j},$ and α, A is able to obtain s_j' from $H(u_j, \alpha^{s_j+r_j}, s_j', r_j, \alpha)$ by guessing attack.
4. Using the obtained s_j', malicious inside user P_i can forge the $P_j's$ response message $H(s_j', K', u_1, \dots, u_t, r_1, \dots, r_t, \alpha)$.

Method 2:

Let $U_a \in$U be an authorized user for a session (K_1), s_a' be his long-term secret, $U_{(k_1)} \subseteq$ U be the qualified set of participants of the session, $(\alpha^{s_{i(k_1)}+r_{i(k_1)}})_{U_i \in U_{(k_1)}}$ be the values that were broadcasted as part of $(M_i)_{U_i \in U_{(k_1)}}$ in step 4 , and $K_{(k_1)} = \alpha^{S_{(k_1)}}$ be the session key.

1. The participant U_a is qualified to determine (k_1) session key as: $K_{(k_1)} = \alpha^{s_a'} \cdot \alpha^{s_{a(k_1)}+r_{a(k_1)}} / \alpha^{r_{a(k_1)}}$.
2. Since $\alpha^{s_{i(k_1)}+r_{i(k_1)}}$ and $r_{i(k_1)}$ are public, he is able to compute $\alpha^{s_i'}$, for all $U_i \in U_{(k_1)}$: $\alpha^{s_i'} = K_{(k_1)} \cdot \alpha^{r_{i(k_1)}} / \alpha^{s_{i(k_1)}+r_{i(k_1)}}$.
3. Suppose that U_a is unauthorized to recover (k_2) session key, $(k_2) \neq (k_1)$. But, he can eavesdrop the exchanged messages, then he is able to compute $\alpha^{s_{j(k_2)}} = \alpha^{s_{j(k_2)}+r_{j(k_2)}} / \alpha^{r_{j(k_2)}}$ for all $U_j \in U_{(k_2)}$, where $U_{(k_2)} \subseteq$U is the qualified set of parties of the session (k_2).
4. The inside adversary U_a can find the key $K_{(k_2)}$ of the session (K_2) as: $K_{(k_2)} = \alpha^{s_b'} \cdot \alpha^{s_{b(k_2)}} = \alpha^{s_b'+s_{b(k_2)}}$ where $U_b \in U_{(k_1)} \cap U_{(k_2)}$. Thus, an insider is able to compute any session key under the assumption that at least one authorized participant for both sessions exists.
5. Then, the inside adversary is able to obtain others' secret shares s_j' $(i = 1, \dots, t)$.
6. Since U_a knows $(h, u_j, \alpha^{s_j+r_j}, s_j', r_j, \alpha)$, U_a is able to forge a $P_j's$ response message $h_j = H(u_j, \alpha^{s_j+r_j}, s_j', r_j, \alpha)$. Thereafter, U_a is able to impersonate P_j.

Unlike to the Olimid's claim, his improved group key transfer protocol is still vulnerable to insider attacks.

3.3 Known Key Security

Suppose an adversary owns a session key $K_{(k_1)}$. We also assume that he had previously eavesdropped values $r_{i(k_1)}$ in step 3, $\alpha^{s_{i(k_1)}+r_{i(k_1)}}$ and α from the broadcasted message in step 4 of session (k_1), then he is able to compute $\alpha^{s_{i(k_1)}} =$

$\alpha^{s_{i(k_1)}+r_{i(k_1)}}/\alpha^{r_{i(k_1)}}$ for all $U_i \in U_{(k_1)}$. Because the session key $K_{(k_1)}$ is exposed, he can also compute the long term secret $\alpha^{s_i'}$, for all $U_i \in U_{(k_1)}$: $\alpha^{s_i'} = K_{(k_1)}/\alpha^{s_{i(k_1)}}$.

Let (k_2) be any previous or future session that has at least one common qualified participant U_b with (k_1), i.e. $U_b \in U_{(k_1)} \cap U_{(k_2)}$. As before, the adversary eavesdropped $r_{b(k_2)}$, $\alpha^{s_{b(k_2)}+r_{b(k_2)}}$, α and computed $\alpha^{s_{b(k_2)}} = \alpha^{s_{b(k_2)}+r_{b(k_2)}}/\alpha^{r_{b(k_2)}}$.

The adversary can now recover the key $K_{(k_2)}$:

$$K_{(k_2)} = \alpha^{s_b'} \cdot \alpha^{s_{b(k_2)}} = \alpha^{s_b'+s_{b(k_2)}}.$$

Therefore, an adversary is able to disclose any session key under the assumption that a session key has been compromised.

4 Conclusion

In 2013, Olimid proposed an improved group key transfer protocol based on a special secret sharing scheme [9]. He claimed that his improved protocol eliminated insider attack and provided known key security. However, our analysis shows that any inside/ outside adversary can obtain the session key and impersonate legal users. Therefore, the improved protocol does not meet the fundamental security goal. Future work could be undertaken to remedy Sun et al. [8] and Olimid protocols.

References

1. Shamir, A.: How to share secret. Communications of the ACM 22(11), 612–613 (1979)
2. Katz, J., Yung, M.: Scalable protocols for authenticated group key exchange. In: Boneh, D. (ed.) CRYPTO 2003. LNCS, vol. 2729, pp. 110–125. Springer, Heidelberg (2003)
3. Nam, J., Paik, J., Kim, U.M., Won, D.: Resource-aware protocol for authenticated group key exchange in integrated wired and wireless networks. Journal of Information Sciences 177, 5441–5467 (2007)
4. Hajyvahabzadeh, M., Eidkhani, E., Mortazavi, S.A., Pour, A.N.: A new group key management protocol using code for key calculation: CKC. Information Science and Applications, 1–6 (2010)
5. Harn, L., Lin, C.: Authenticated group key transferprotocol based on secret sharing. IEEE Transactions on Computers 59(6), 842–846 (2010)
6. Nam, J., Paik, J., Won, D.: A security weakness in Abdalla et al.'s generic construction of a group key exchange protocol. Journal of Information Sciences 181(1), 234–238 (2011)
7. Nam, J., Kim, M., Paik, J., Won, D.: Security Weaknesses in Harn-Lin and Dutta-Barua protocols for group key establishment. KSII Transactions on Internet and Information Systems 6(2), 751–765 (2012)
8. Sun, Y., Wen, Q., Sun, H., Li, W., Jin, Z., Zhang, H.: An authenticated group key transferprotocol based on secret sharing. Procedia Engineering 9, 403–408 (2012)
9. Olimid, R.F.: On the security of an authenticated group key transfer protocol based on secret sharing. In: Mustofa, K., Neuhold, E.J., Tjoa, A.M., Weippl, E., You, I. (eds.) ICT-EurAsia 2013. LNCS, vol. 7804, pp. 399–408. Springer, Heidelberg (2013)

Analysis of Cyber Attacks and Security Intelligence*

Youngsoo Kim[1], Ikkyun Kim[1], and Namje Park[2,**]

[1] Cyber Security Research Laboratory,
Electronics and Telecommunications Research Institute (ETRI),
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea
{yskim,ikkim}@etri.re.kr
[2] Department of Computer Education, Teachers College,
Jeju National University, Jeju, Korea
namjepark@jejunu.ac.kr

Abstract. A cyber attack is deliberate exploitation of computer systems, technology-dependent enterprises and networks. Cyber attacks use malicious code to alter computer code, logic or data, resulting in disruptive consequences that can compromise data and lead to cybercrimes, such as information and identity theft. Cyber attack is also known as a computer network attack (CNA). Cyber attacks occurred targeting banks and broadcasting companies in South Korea on March 20. The malware involved in these attacks brought down multiple websites and interrupted bank transactions by overwriting the Master Boot Record (MBR) and all the logical drives on the infected servers rendering them unusable. It was reported that 32,000 computers had been damaged and the exact amount of the financial damage has not yet been calculated. More serious is that we are likely to have greater damages in case of occurring additional attacks, since exact analysis of cause is not done yet. APT(Advanced Persistent Threat), which is becoming a big issue due to this attack, is not a brand new way of attacking, but a kind of keyword standing for a trend of recent cyber attacks. In this paper, we show some examples and features of recent cyber attacks and describe phases of them. Finally, we conclude that only the concept of security intelligence can defend these cyber threats.

Keywords: Cyber Attacks, Security Intelligence, MBR, APT, Threat.

1 Introduction

Advanced persistent threat (APT) usually refers to a group, such as a foreign government, with both the capability and the intent to persistently and effectively target a specific entity. The term is commonly used to refer to cyber threats, in particular that of Internet-enabled espionage using a variety of intelligence gathering techniques to access sensitive information, [1] but applies equally to other threats

* This research was funded by the MSIP(Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_73,

such as that of traditional espionage or attack[2]. Other recognized attack vectors include infected media, supply chain compromise, and social engineering. Individuals, such as an individual hacker, are not usually referred to as an APT as they rarely have the resources to be both advanced and persistent even if they are intent on gaining access to, or attacking, a specific target[3].

APT(Advanced Persistent Threat), which is becoming a big issue due to this attack, is not a brand new way of attacking, but a kind of keyword standing for a trend of recent cyber attacks[1]. Originally, this word was used as a type of specific security threats in US Air Force, but it has been recognized as a new paragon of cyber attacks since stuxnet, which is a malware being used to hack Iran's nuclear power facilities, was found. Stuxnet greatly impacted the society to show the possibility that cyber threats, regarded as one of the group activities for political end before, can paralyze industry control system and cause a large-scale of financial damages.

After stuxnet, similar cyber attacks have occurred all over the world. In 2009, Operation Aurora being used to leak secrets and falsify source codes from more than 30 huge firms such as Google, Adobe, and Juniper, became a diplomatic issue between US and China, since, US claimed that Chinese hackers took lead attacks using this malware, Operation Aurora, but China denied and criticized US. Hacking groups such as Anonymous or LulzSec claiming to stand for Hacktivism have attacked HBGary, a security company, US agencies of FBI and CIA, and subsidiaries of SONY. In 2011, The Night Dragon hacking attacks were targeted at some of the world's largest petrochemical companies, including Shell, Exxon Mobile, BP, Marathon Oil, ConocoPhillips, and Baker Hughes. Numerous critical data in gas and oil area were leaked by this attack. Additionally, EMC/RSA, a security company, was attacked by cyber terror using social engineering methods and classified data of SecureID, OTP(One-Time Password) solution, was stolen. In March of the same year, more than 150 French diplomats' computers were attacked and Paris G20 files were stolen in cyber attack. Furthermore, around 760 firms, including almost 20 percent of the top Fortune 100 companies in the US, turned out to have suffered similar cyber attack, under investigation of hacking case which Lockheed Martin, America's largest defense contractor, was attacked by massive cyber attack.

We can find some domestic examples of cyber attack at ease. 7.7 DDoS attack and 3.4 DDoS attack occurred at July of 2009 and March of 2011 targeting popular web pages and portals, and caused economic losses of 0.4 million dollars. NH's online network paralysis with long-term penetration had become a big issue, because NH's banking services delayed for a few days, and had given rise to need of forensic readiness[2]. Additionally, in 2011, a cyber terror thought to be jamming of North Korea had caused paralysis of wireless networks in northern part of Seoul for a while, and we had become to know that North Korea's cyber offensive skills are as good as or better than their counterparts. Actually, a reporter of Fox news said that North Korean military has around 30,000 electronic warfare specialists and they have become the elite core of the military. He also said that the regime now culls the brightest students from the nation's universities and funnels them into special secret schools that concentrate on hacking and developing cyber warfare programs, and North Korea has the capability to paralyze the US Pacific Command and cause extensive damage to defense networks inside the US.

Features of recent cyber attacks are as follows. First, it penetrates an exact target and steals critical data using intelligent security threats like zero-day attacks or rootkits. A zero-day attack or threat is an attack that exploits a previously unknown vulnerability in a computer application, meaning that the attack occurs on "day zero" of awareness of the vulnerability. This means that the developers have had zero days to address and patch the vulnerability. Zero-day exploits (actual software that uses a security hole to carry out an attack) are used or shared by attackers before the developer of the target software knows about the vulnerability[6]. A rootkit is a stealthy type of software, often malicious, designed to hide the existence of certain processes or programs from normal methods of detection and enable continued privileged access to a computer. The term rootkit is a concatenation of "root" (the traditional name of the privileged account on UNIX OS) and the word "kit" (which refers to the software components that implement the tool). The term "rootkit" has negative connotations through its association with malware[4]. Second, it makes progress in the long term patiently and stealthily not to be discovered by target's defense system. The time required to complete mission can be a couple of months or years, but it could be longer if we consider periods of making plans and preparing attacks for achieving goals. Finally, it is performed having goals with serious impact of stealing critical information related militaries, politics, economies, etc., or paralyzing industry control systems, not simple goals like showing hacking powers or leaking someone's banking accounts. In above instances, it is highly likely that well-financed groups or some countries' intelligence services are involved.

2 Typical Phases of Cyber Attacks

Recent cyber attacks have the following phases to achieve goals: Reconnaissance, Preparation, Targeting, Further Access, Data Gathering, and Maintenance[5].

Reconnaissance is a phase of passively gathering information about their target to identify the best targeting method. This may include research into the location of the target's offices, the location of their computers, technologies used by the company, how they communicate (between offices, with customers, suppliers and shareholders), their employees, their employees' contact details, interests and contacts.

Preparation is a step of developing and testing appropriate tools and techniques to target their intended victim. This may include scanning to determine vulnerabilities, writing malicious code or acquiring code, drafting socially engineered emails, determining which email account to send socially engineered emails from, acquiring necessary hardware (such as USB flash drives), determining what infrastructure to use to launch the attack and for command and control communications, registering for and setting up necessary accounts (email addresses, callback domains etc.) and conducting testing.

In targeting step, the attacker launches their attack and monitors for signs of compromise or failure. The sender may attempt to connect remotely to a server to exploit a vulnerability, strategically place a USB flash drive or give one to a target, send socially engineered emails and if possible, check for bounce back notifications, monitor command and control infrastructure for beaconing activity from the victim, try to connect inbound to the potentially compromised computer, or await feedback from an insider.

Once an attacker has successfully gained access to a computer network they will usually need a phase of further access to try to identify where in the network they are and move laterally within the network to access data of interest and to install additional backdoors. This will usually require a return to step 2 (Preparation) and step 3 (Targeting), the upload of tools and malicious software, privilege escalation, network enumeration and identification of vulnerable hosts on which to install backdoors. It may also involve gaining access to the domain controller to obtain password hashes, covering tracks by altering logs, and accessing mail or file servers to enable data gathering. Once an attacker has identified information of interest they will try to gather this information and exfiltrate it. They may do this using a 'smash and grab' approach, trying to exfiltrate the desired data before it is detected, or they may opt for a 'low and slow' approach in which they exfiltrate the data in small quantities over a longer period. Once an attacker has gained access to a network for information gathering purposes they will usually attempt to maintain their access. This may involve minimizing the amount of malicious activity they generate on the network to avoid detection, periodically communicating with backdoors on the network to ensure they are working as intended, and making changes as appropriate. If automated data gathering tools are in use, it may also involve modifying search terms or the exfiltration path, volume or frequency.

Maintenance also requires maintaining callback domains and any intermediary infrastructure used to communicate with the backdoors. If access is lost, the attacker may return to step 1 (Reconnaissance) or step 2 (Preparation) in an attempt to regain access.

3 Security Intelligence

Since it could hardly defense recent diverse cyber threats using conventional platform-based analyzing technologies like IDS/IPS, there is a need for a newly defending way, a concept of Security Intelligence. It is an advanced analysis technology for security information, improving security intelligence by analyzing relations of data and security events being generated from networks, systems, and applications of IT infrastructures, in order to respond unknown fatal attacks including APT. Gartner group, an information technology research and advisory company providing technology related insight, has a concept that Security Intelligence is a methodology having interaction among various security technologies and defines it as context-based analysis technology combining information from diverse sources and having relations among them[6]. Thus, they assess the notion of Security Intelligence is presented as a form of context-aware security in the short term, and it will be continued for about 10 years. Additionally, they prospect that it is expected to be a spying out method for unknown attacks by analyzing relations of diverse characteristic factors of network, outside the pale of pattern-based threat control methods being used in conventional security products.

4 Conclusion

Cyber attacks occurred targeting banks and broadcasting companies in South Korea on March 20. The malware involved in these attacks brought down multiple websites and interrupted bank transactions by overwriting the Master Boot Record (MBR) and all the logical drives on the infected servers rendering them unusable. It was reported that 32,000 computers had been damaged and the exact amount of the financial damage has not yet been calculated. More serious is that we are likely to have greater damages in case of occurring additional attacks, since exact analysis of cause is not done yet. APT(Advanced Persistent Threat), which is becoming a big issue due to this attack, is not a brand new way of attacking, but a kind of keyword standing for a trend of recent cyber attacks. In this paper, we show some examples and features of recent cyber attacks and describe phases of them. Finally, we conclude that only the concept of security intelligence can defend these cyber threats.

Acknowledgments. This paper is extended from 2013 5th International Conference on Data Mining and Intelligent Information Technology Applications (June, 2013). This research was funded by the MSIP(Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013. The corresponding author is Namje Park.

References

1. Advanced Persistent Threat (APT), http://en.wikipedia.org/wiki/Advanced_persistent_threat
2. Pangalos, G., et al.: The Importance of Corporate Forensic Readiness in the information security framework. In: 2010 Workshops on Enabling Technologies (2010)
3. Zero-day attack, https://en.wikipedia.org/wiki/Zero-day_attack
4. Rootkit, http://en.wikipedia.org/wiki/Rootkit
5. Rivner, U.: Anatomy of an Attack, http://blogs.rsa.com/rivner/anatomy-of-an-attack/
6. MacDonald, N.: The future of information Security is Context Aware and Adaptive. Gartner
7. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
8. Park, N.: Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
9. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
10. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)

11. Park, N., Kim, Y.: Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 193–202. Springer, Heidelberg (2010)
12. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 142–152. Springer, Heidelberg (2010)
13. Park, N.: Customized Healthcare Infrastructure Using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
14. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
15. Kim, Y., Park, N., Hong, D.: Enterprise Data Loss Prevention System Having a Function of Coping with Civil Suits. In: Lee, R. (ed.) Computers,Networks, Systems, and Industrial Engineering 2011. SCI, vol. 365, pp. 201–208. Springer, Heidelberg (2011)
16. Kim, Y., Park, N., Won, D.: Privacy-Enhanced Adult Certification Method for MultimediaContents on Mobile RFID Environments. In: Proc. of IEEE International Symposium onConsumer Electronics, pp. 1–4. IEEE, Los Alamitos (2007)
17. Kim, Y., Park, N., Hong, D., Won, D.: Adult Certification System on Mobile RFID ServiceEnvironments. Journal of Korea Contents Association 9(1), 131–138 (2009)
18. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
19. Park, N.: The Implementation of Open Embedded S/W Platform for Secure Mobile RFID Reader. The Journal of Korea Information and Communications Society 35(5), 785–793 (2010)
20. Park, N.: Mobile RFID/NFC Linkage Based on UHF/HF Dual Band's Integration in U-Sensor Network Era. In: Park, J.J. (J.H.), Kim, J., Zou, D., Lee, Y.S. (eds.) ITCS & STA 2012. LNEE, vol. 180, pp. 265–271. Springer, Heidelberg (2012)
21. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)

Protection Profile for PoS (Point of Sale) System*

Hyun-Jung Lee[1], Youngsook Lee[2], and Dongho Won[1,**]

[1] School of Information and Communication Engineering, Sungkyunkwan University, Korea
hjlee@kosyas.com, dhwon@security.re.kr
[2] Department of Cyber Investigation Police, Howon University, Korea
ysooklee@howon.ac.kr

Abstract. A PoS system immediately obtains the data related to the sale at the time and place of purchase. It provides an initial interface for the credit card transaction to happen. Due to its dealing with sensitive data such as credit card information, many relevant organizations have been trying to suggest security standards. However, there still is no PoS system that guarantees security, which results in a lot of hacked PoS systems in different countries. This paper intends to draw out security functional requirements for a PoS system based on the CC, which can be used as a reference for its security evaluation.

Keywords: Protection Profile, CC(Common Criteria), PoS(Point of Sale), Security Requirement, Vulnerability.

1 Introduction

In April 2010, the police identified an international hacker group that hacked into the PoS systems for a credit card, which are widely deployed in restaurants and gas stations, and leaked the customers' credit card data abroad so that they were re-leaked to 49 countries and used to make counterfeit cards. The police reports said that there have been accidents related to a PoS system, where credit card data stored in the PoS systems in restaurants and gas stations were leaked by hacking or secret information stored in the magnetic of a credit card was automatically leaked to an email at the time of payment by a malicious program. More than 100,000 credit cards' data have been leaked outside the country since the second half of 2009. Leaked credit card data cases keep being reported since 2007, with damages to dozens of stores so far. The aspect of data leak used to be simple as direct drain of database containing the credit card information but now has evolved into real-time leak through an email at the point of payment. Though several organizations are trying to draw out a guideline for improvement of PoS systems, most stores are still using the systems with vulnerabilities. Since researches on developing a PP for PoS systems have been limited to a PoS

* This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0301-13-3007) supervised by the NIPA(National IT Industry Promotion Agency).
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_74,

terminal, which is only a part of a PoS system[1], there is no guidance on dealing with vulnerabilities of the other components.

This paper draws out basic security functions that a PoS system should have from the Common Criteria (CC) to suggest a protection profile, which will be used as guidance in developing and evaluating a PoS system and hopefully help reduce crimes that occur through the system. The rest of the paper is organized as follows: In Section 2, we review related works. In Section 3, we propose the protection profile for PoS System. In Section 4, we present our conclusions.

2 Related Work

2.1 CC and Protection Profile

The Common Criteria for Information Technology Security Evaluation (Common Criteria or CC) is an international standard (ISO/IEC 15408) established with the objectives to integrate various evaluation standards from different countries and allow mutual recognition of the evaluation results between the countries that agrees on the idea[2][3][4]. The CC provides a special construct called Protection Profile (PP). Whereas an ST always describes a specific TOE, a PP is intended to describe a TOE type (e.g. firewalls). The same PP may therefore be used as a template for many different STs to be used in different evaluations. A PP must contain a PP introduction, conformance claim, security problem definition, security objectives, extended components definition, and security requirements. [2][3][4]

3 Security Problem Definition

This section defines the security problem. The security problem definition consists of three subsections, threats, organizational security policies and assumptions. The process of deriving the security problem definition falls outside the scope of the CC[2].

Threats. This subsection of the security problem definition shows the threats that are to be countered by the TOE. A threat consist of a threat agent, an asset and an adverse action of that threat agent on that asset[2]. The specification of threats should include all threats detected up to now, if it is not done the TOE may provide inadequate protection. In other words, if the specification of threats is insufficiency, the assets may be exposed to an unacceptable level of risk. So we reviewed most related papers and reports ([5,6,7,8]). In the result, we derive the threats in Table 1.

Organizational Security Policies(OSP). The organizational security policies define rules and policies of organization, supporting the TOE security, which operates the TOE. The organizational security policies for this paper are described in Table 1.

Assumptions. The assumptions are "givens" regarding secure usage of the TOE and necessary conditions in order to guarantee completeness of the TOE security, because the TOE cannot support all security functions. So we reviewed most related papers and reports ([5,6,7,8]). In the result, we derive the minimum assumptions for this protection profile in Table 1

Table 1. Security Problem Definition

THREAT	DESCRIPTION
T.Stored Data Disclosure	A threat agent may disclose, modify, or delete the TSF data stored in the TOE(e.g. credit card information, authentication data, etc.) in an unauthorized way.
T.Overgenerated Log Record	A threat agent may get information from excessively generated log records of the TOE.(PoS Terminal Only)
T.Undergenerated Log Record	TSF data may be exposed during the operation of the TOE due to a function that does not generate a log. (PoS Terminal Only)
T.Recovery	A threat agent may try to recover deleted TST data of the TOE using a commercial tool.
T.Transferred Data Disclosure	A threat agent may disclose, modify, or delete the TSF data being transferred from/to the TOE in an unauthorized way.
T.Unauthorized Access	A threat agent may access the TOE by masquerading as an authorized administrator.
T.Consecutive Authentication Attempt	A threat agent may access the TOE by consecutively attempting authentication.
T.TOE Access NW	A threat agent may cause outflow, deletion, or fabrication of the user data and TSF data stored in the TOE through network.
T.System Compromise	A threat agent may interrupt the execution of the TOE by launching an attack against the TOE or its operational environment.
OSP	**DESCRIPTION**
P.Data Policy Setup	In case of collecting and storing card data, establish a policy on the storage and destruction of the data, based on which to restrict the capacity and period of storage in business, law, and regulations.
P.PCIDSS	A TOE administrator shall follow the PCIDSS policy.
P.Data Policy Setup	In case of collecting and storing card data, establish a policy on the storage and destruction of the data, based on which to restrict the capacity and period of storage in business, law, and regulations.
P.Audit	To trace responsibilities on all security-related activities, all security-related events shall be recorded; the recorded data shall be maintained and reviewed.
P.PCIDSS	A TOE administrator shall follow the PCIDSS policy.
ASSUMPTION	**DESCRIPTION**
A.NW Access Control	In order to protect the internal network where the TOE is installed, a policy to protect wired/wireless network will be established, which requires installation and maintenance of a security system such as firewall. The security system will cut off traffic that does not comply with specified security policies and control access to the PoS and branch server.
A.Physical Security	A branch server, which is a part of the TOE, will be located in a physically secure environment that can only be accessed by an authorized administrator. A PoS terminal, which is also a part of the TOE, will be fixed to a place.
A.Trusted Admin	An authorized administrator of the TOE is not malicious, receives proper training on the TOE management functions, and follows the administrator guidelines.
A.OS Enhancement	Services or means not required by the TOE will be removed from the operating system and vulnerabilities of the operating system will be fixed properly to ensure its reliability and stability.

Table 1. *(continued)*

NAME	DESCRIPTION
A.Default Change	The default values provided by a vendor (e.g. default password, settings, etc. at the time of release of the TOE) will be changed according to the level of security before the installation is complete.
A.Virus Vaccine SW	A system on which the TOE is installed shall have vaccine software to block vulnerabilities and malicious viruses inflowing through the network.
A.Minimal Info Disclosure	User's manual of a PoS terminal does not provide too much information, such as its default password, how to reset the terminal, etc.
A.Security Maintenance	When the internal network environment changes due to change in the network configuration, host increase/decrease, service increase/decrease, etc., the changed environment and security policy will immediately be reflected in the TOE operational policy so that security level can be maintained to be the same as before.

4 Security Objectives

Security objectives are intended solution to the problem defined by the security problem definition. They can be categorized into security objectives for the TOE and for the operational environment. The former means security functionality that the TOE provides to solve a certain part of the problem, which consists of a set of objectives that the TOE should achieve. The latter means technical and procedural measures to assist the TOE in correctly providing its security functionality, which consists of a set of statements describing the goals that the operational environment should achieve [9,10]. The Security Objectives for this paper are described in Table 2.

Table 2. Objective

NAME	DESCRIPTION
O.Stored Data Protection	The TOE shall protect the TSF data stored in it from unauthorized disclosure, modification, or deletion. A PoS terminal provides various means of storage, such as static RAM, CF memory, or Hard Disk Drive, for the TSF data (e.g. credit card information). The TOE shall be able to respond to a threat to the TSF data regardless of the type of storage.
O.Audit	The TOE shall correctly generate and securely maintain the record of all security-relevant events to ensure they can be traced and shall provide a means to review the records. It shall also provide a function to deal with audit data storage exhaustion.
O.Management	The TOE shall provide a means for the authorized administrator of the TOE to efficiently manage the TOE in a secure manner.
O.IA	The TOE shall identify and authenticate an administrator that attempts to access the data stored in the PoS terminal. It shall detect and deal with consecutive authentication failures.
O.TOE Operation	It is guaranteed that the TOE always functions normally.
O. Deletion	TSF data shall be deleted in a way that it cannot be recovered.
O.Access Control Policy	The user data and TSF data stored in the TOE shall be protected from outflow, deletion, or fabrication through network.

Table 2. *(continued)*

NAME	DESCRIPTION
O.Transferred Data Protection	All TSF data being transferred between PoS systems and those transferred for the sake of security management shall be protected from unauthorized disclosure and modification.
OE.Trusted Administrator	An authorized administrator of the TOE is not malicious, receives proper training on the TOE management functions, and follows the administrator guidelines and protocol.
OE.Secure Management	The TOE shall be delivered and installed in a secure manner. An authorized administrator of the TOE and authorized user in the operational environment shall configure and manage the TOE in a secure manner.
OE.OS Enhancement	An authorized administrator of the TOE and authorized user in the operational environment shall fix the vulnerabilities of the OS so that the TOE can be tested for its accurate operation and there will be no interference between the TOE and other applications.
OE.NW Access Control	In order to protect the internal network where the TOE is installed, a policy to protect wired/wireless network shall be established, which requires installation and maintenance of a security system such as firewall. The security system shall cut off traffic that does not comply with specified security policies and control access to the PoS and branch server.
OE.Security Maintenance	When the internal network environment changes due to change in the network configuration, host increase/decrease, service increase/decrease, etc., the changed environment and security policy shall immediately be reflected in the TOE operational policy so that security level can be maintained to be the same as before.
OE.Default Value Change	The default values provided by a vendor (e.g. default password, settings, etc. at the time of release of the TOE) shall be changed according to the level of security before the installation is complete.
OE.PCIDSS	The TOE administrator shall follow the PCIDSS policy.
OE.Virus Vaccine SW	A system on which the TOE is installed shall have vaccine software to block vulnerabilities and malicious viruses inflowing through the network.
OE.Physical Security	The TOE shall be located in a physically secure environment and protected from unauthorized access.
OE.Data Policy Setup	In case of collecting and storing card data, establish a policy on the storage and destruction of the data, based on which to restrict the capacity and period of storage in business, law, and regulations.

5 Security Requirement

Security requirements are intended to satisfy the security objectives. The TOE should be developed to satisfy the security objectives. Table 3 shows the mandatory security functional requirements for a PoS system.

Table 3. Security functional component

Security functional class	Security functional component
Security audit	FAU_ARP.1, FAU_GEN.1, FAU_GEN.2, FAU_SAA.1, FAU_SAR.1, FAU_STG.1, FAU_STG.3, FAU_STG.4

Table 3. *(continued)*

Security functional class	Security functional component
Cryptographic support	FCS_CKM.1, FCS_CKM.2, FCS_CKM.4, FCS_COP.1
User data protection	FDP_IFC.1, FDP_IFF.1
Identification and authentication	FIA_AFL.1, FIA_ATD.1, FIA_SOS.1, FIA_UAU.2, FIA_UID.2
Security management	FMT_MOF.1, FMT_MTD.1, FMT_MTD.2, FMT_SMF.1, FMT_SMR.1
Protection of the TSF	FPT_STM.1, FPT_TST.1, FTP_ITC.1, FTP_ITT.1
TOE access	FTA_SSL.3

6 Conclusions

Unlike a credit card payment system, this connects to the telephone line, a PoS system used in department stores, large stores, shops, franchise restaurants, and gas stations processes payment and management through the Internet. More and more people say that security should be considered for a PoS system, yet accidents resulting from its vulnerabilities keep increasing. This paper derives mandatory security functional requirements based on the CC (ISO/IEC 15408) in order to make an objective evaluation of the security of a PoS system possible. It also intends to be used in introducing or developing a PoS system, eventually in improving its security.

References

1. Pedersen, A., Hedegaard, A.: Designing a Secure Point-of-Sale System. In: Proceedings of the Fourth IEEE International Workshop on Information Assurance (IWIA 2006) (2006)
2. Common Criteria, Common Criteria for Information Technology Security Evaluation; part 1: Introduction and general model, Version 3.1 R1, CCMB-2006-09-001 (September 2006)
3. Common Criteria, Common Criteria for Information Technology Security Evaluation; part 2: Security functional components, Version 3.1 R2, CCMB-2007-09-002 (September 2007)
4. Common Criteria, Common Criteria for Information Technology Security Evaluation; part 3: Security assurance components, Version 3.1 R2, CCMB-2007-09-003 (September 2007)
5. PCI Security Standards Council "Payment Card Industry Data Security Standard Version 1.1", Release 09.07.2006
6. SPVA "End-to-End Encryption Security Requirements Revision 1.0" (May 27) (June 25, 2010)
7. Dr. Neal Krawets Hacker Factor Solutions "PoS Vulnerability Version 2.0" (August 27, 2007)
8. PNC SAC, "Security Requirements For an EFTPoS Terminal Version 1.2.1" (March 16, 2006)
9. Lee, H.-J., Won, D.: Protection Profile for Data Leakage Protection System. In: Kim, T.-H., Adeli, H., Slezak, D., Sandnes, F.E., Song, X., Chung, K.-I., Arnett, K.P. (eds.) FGIT 2011. LNCS, vol. 7105, pp. 316–326. Springer, Heidelberg (2011)
10. Lee, H.-J., Won, D.: Protection Profile for Personal Information Security System. In: IEEE TrustCom 2011, pp. 806–811 (2011)

A Probabilistic Timing Constraint Modeling and Functional Validation Approach to Dynamic Service Composition for LBS

Weimin Li[1], Xiaohua Zhao[1], Jiulei Jiang[2], Xiaokang Zhou[3], and Qun Jin[4,3]

[1] School of Computer Engineering and Technology, Shanghai University, Shanghai, China
[2] School of Computer, Beifang University of Nationalities, Yinchuan, Ningxia, China
[3] Graduate School of Human Sciences, Waseda University, Tokorozawa, Japan
[4] College of Information Engineering, China Jiliang University, Hanzhou, China
wmli@shu.edu.cn, {flowerencedew,wjjl2005}@126.com,
xkzhou@ruri.waseda.jp, jin@ waseda.jp

Abstract. Location Based Services (LBS) is a kind of real-time service with uncertain factors, and its modeling and validation is essential. In this paper, with the introduction of probability, we propose a Color Probability-TCPN (CP-TCPN) by using the tokens with specific colors as the research objects and redefining several relative parameters. We use CP-TCPN to realize modeling and functional verification of the dynamic services composition for LBS. Simulation result is presented to illustrate the application of CP-TCPN in the modeling and analyzing of the real-time system with uncertain factors.

Keywords: CP-TCPN, Probability, Colors, Service composition, Modeling, Functional validation.

1 Introduction

LBS is a kind of value-added service that provides services to the users based on their geographic location. Nowadays, with the development of the mobile communication technology, the positioning technology and the mobile terminal technology, LBS has become an indispensable part of people's daily life. And the research on the modeling and function validation for LBS dynamic services composition is becoming a hotspot. There have been lots of modeling and functional validation methods, including Petri Net and the extension methods based on it, such as PPN (Probabilistic Petri Net) [1], SPN (Stochastic Petri Net) [2] and TCPN (Timing Constraint Petri Net) [3]. TCPN was first proposed by Tsai et al. [3]. They described the definition, modeling method, analysis method and its usage in the real-time system. Lin et al. introduced timing constraint into the control process of workflow, validated and analyzed the model with the relative theory of TCPN [4]. In order to solve the problem that Time Petri Net and Timed Petri Net cannot model and analyze the project performance properly, Yu et al. used TCPN to model and analyze the Project Performance [5]. For real-time systems' modeling and analysis, Tsai et al. used TCPN to model the systems and analyze the models'

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_75,

schedulability, and demonstrated the reliability through a real-time system [6]. In addition, TCPN was also used for verification of temporal consistency in Web service process [7]. Tsai et al. proceeded a further study on the verification of distributed real-time system, and proposed both static and dynamic analysis procedures to verify timing properties of distributed real-time systems [8]. However, these methods cannot be used in real-time systems' modeling and functional validation as a whole.

In this paper, by introducing the probability factors into TCPN and redefining the relevant parameters based on PPN, we proposed a new modeling method named Color Probability-TCPN (CP-TCPN) which can meet the needs of modeling and functional validation for LBS dynamic services composition. In other words, CP-TCPN is proposed to model and analyze uncertain real-time systems. By using CP-TCPN, the system's stability could be quantitatively analyzed, and the method can also quantitatively evaluate the service composition scheme and calculate the system's time complexity comprehensively.

2 Color Probability Timing Constraint Petri Nets

TCPN is six-tuple which was originally proposed by Tsai in 1995 on the basis of Time Petri Net and Timed Petri Net. As the probability isn't considered, TCPN cannot describe the real-time systems with uncertain factors, such as LBS. And it also cannot quantitatively analyze the probability of the system. In this paper, we propose Color Probability-Timing Constraint Petri Nets by introducing probability into TCPN.

Definition 1. *CP-TCPN* is eight-tuple through adding the *token*'s color and probability, and the definition is given as follows:

$$CP\text{-}TCPN=< P, T, F, C, Pro, T_C, D, M >$$

where $T_C= \{T_{min}, T_{max}\}$ is the timing constraint interval of places and transitions. In this paper, the bound value of the intervals can be adjusted dynamically with the change of requests conditions; $P= \{p_1, p_2, \ldots\ldots, p_n\}$ is the set of places. In the sub-vector $p_i=<T_C, C, Pro>$, T_C is the timing constraint of p_i, C represents the color set of p_i, and Pro is the usage probability of p_i. If p_i is an output place, the probability is set to 1; $T= \{t_1, t_2, \ldots\ldots, t_m\}$ is the set of transitions; $F= \{F_1, F_2, \ldots\ldots, F_k\}$ is the set of arcs that connect the places with transitions. In the sub-vector $F_j=<P_I / P_O, Pro>$, P_I represents the input place set of arc F_j, and P_O is the set of output places. Pro is the usage probability of F_j, and the value of Pro is different along with different firing condition and tokens' usage; $C= \{c_1, c_2, \ldots\ldots, c_n\}$ is the set of colors. In sub-vector $c_i=<cond, color, F_c, T_{arr}(c_i)>$: $cond$ represents the usage conditions of the tokens belonging to the input places with certain color, $color$ represents the token t_i's color, F_c is the set of the t_i's corresponding arcs, and $T_{arr}(t_i)$ represents the minimum elapsed time before the arrival of token t_i with its color c_i. We explain C with a simple example as shown in Fig. 1.

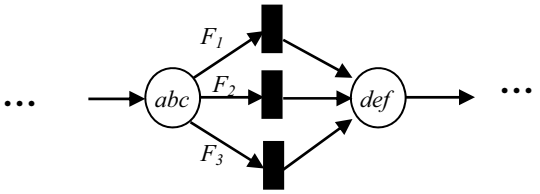


Fig. 1. An example of C (color) in CP-TCPN

C= {a, b, c, d, e, f } ; $color(t_1)=a$, $color(t_2)=b$, $color(t_3)=c$;

c_1=<$cond_1|cond_2$, a, {F_1, F_2}, $T_{arr}(c_1)$> ; c_2=<$cond_2$, b, {F_2, F_3}, $T_{arr}(c_2)$> ;

c_3=<$cond_3$, c, {F_1, F_2, F_3}, $T_{arr}(c_3)$> ;

And, *Pro* is the probability set of the model and the model's parameters, such as the conditions' fire probability, the selection probability of tokens and arcs, the schedule probability of transitions and places, and so on. D is the set of the transitions' firing duration. And the transition firing duration D is redefined as follows.

i. In this paper, D is a three-tuple D=<D_{min}, D_{max}, S>:

D_{min} and D_{max} represent the minimum and maximum firing duration; S represents the probability distribution of the transitions firing moment. The distribution function S with its distribution interval is (D_{min}, D_{max}), which can be mapped to the standard normal distribution in the interval (-∞,+∞). In this study, we define S by Eq. (1)

$$S(D) = \sqrt{\frac{\pi}{2}} \times \frac{1}{(D_{max}-D_{min})cos^2(\frac{\pi(2D-D_{max}-D_{min})}{2(D_{max}-D_{min})})} \times e^{-\frac{1}{2}tan^2(\frac{\pi(2D-D_{max}-D_{min})}{2(D_{max}-D_{min})})} \quad (1)$$

ii. Supposing the special situation that the firing duration of a transition is smaller than D_{min} or larger than D_{max} will not happen, the firing duration D satisfies Eq. (2).

$$\int_{D_{min}}^{D_{max}} S(D)\, d(D) = 1 \quad (2)$$

Finally, M= {$M(p_1)$, $M(P_2)$, …, $M(P_n)$} is a set of marking about the tokens with n-vector, where the sub-vector $M(P_i)$ represents the amount of tokens in place P_i at a specific moment; and M_0 is the initial state. Fig. 2 shows a simple example of CP-TCPN.

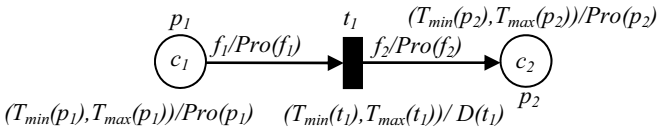


Fig. 2. An example of *CP-TCPN*

3 Probability Analysis Method

Definition 2. Weak schedulable: Analyzing the transition's schedulability without considering the tokens' arriving time or assuming the tokens' arriving time of the model is same, the transition is weak schedulable if and only if the following inequality is established:

$$T_{LFE}(t_j) - T_{EFB}(t_j) \geq D_{max} \tag{3}$$

where $T_{LFE}(t_j)$ and $T_{EFB}(t_j)$ respectively represents the latest fire ending time and the earliest fire beginning time, and their calculating method is given as follows:

$$T_{LFE}(t_j)=min\{min\{\ T_{max}(p_i)\},\ max\ \{T_{min}(p_i)\}+\ T_{max}(t_j)\} \tag{4}$$

$$T_{EFB}(t_j)=max\{T_{min}(p_i)\}+\ T_{min}(t_j) \tag{5}$$

and $p_i \in IP(t_j)(i=1,2,\ldots,n)$.

Definition 3. Strong schedulable: A transition of the model is strong schedulable if and only if the following inequality is established with the consideration of tokens' arriving time:

$$T'_{LFE}(t_j) - T'_{EFB}(t_j) \geq D'_{max} \tag{6}$$

where $T'_{LFE}(t_j)$ and $T'_{EFB}(t_j)$ respectively represents the latest fire ending time and the earliest fire beginning time, and the calculating method is shown as follows;

$$T'_{EFB}(t_j)=max\ \{min\{T_{arr}(c_i)\}+T_{min}(p_i)\}+\ T_{min}(t_j); \tag{7}$$

$$T'_{LFE}(t_j)=min\{min\{max\{T_{arr}(c_i)\}+T_{max}(p_i)\},\ max\{max\{T_{arr}(c_i)\}+T_{min}(p_i)\}+T_{max}(t_j)\}; \tag{8}$$

and $c_i \in p_i$, $p_i \in IP(t_j)(i=1,2,\ldots,n)$.

We define the transition's firing duration as a time interval, so that we can give a quantitative analysis of the transition's strong schedulability. We name the successful firing probability as $P_{FS}(t_i)$, and the computational formula is described as follows:

$$P_{FS}(t_i) = \int_{D_{min}(t_i)}^{\min\{D_{max}(t_i),T_{LFE}(t_i)-T_{EFB}(t_i)\}} S(D)d(D) \tag{9}$$

The model's schedulability analysis is based on the analysis of every transition, and we can achieve strong schedulable probability after the probability calculation of each transition. But for a large scale model, we need to simplify the model during the computing process. Since the generated model only has sequential structure, we can achieve the whole model's reachable probability after necessary calculation.

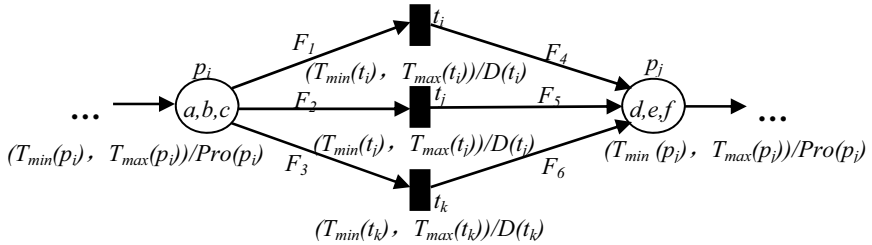


Fig. 3. A model example with probability

Next, we explain how to calculate the services composition scheme generating probability. From the example, we can calculate the services composition scheme generating probability accurately. However, at the same time we also need to gather the users' feedback to adjust the relative parameters' value, because we can achieve a changing value of the services composition scheme generating probability. The model is shown in Fig. 3.

Table 1. The token generation probability

Output token(p_j)	
Token	*Probability*
{d}	$P(d)= 1\times[P_1\times P_1(a)\times P_1(F_1)+ P_2\times P_2(a)\times P_2(F_1)+ P_3\times P_3(c)\times P_4(F_1)]$
{e}	$P(e)=1\times[P_1\times P_1(a)\times P_1(F_2)+ P_2\times[P_2(a)\times P_2(F_2)+ P_2(b)\times P_3(F_2)] + P_3\times P_3(c)\times P_4(F_2)]$
{f}	$P(f)= 1\times[P_2\times P_2(b)\times P_3(F_3)+ P_3\times P_3(c)\times P_4(F_3)]$

$C= \{a, b, c,\ \ d, e, f\}$; $color(c_1)=a$, $color(c_2)=b$, $color(c_3)=c$;

$c_1=<cond_1|cond_2, a, \{F_1, F_2\}, 0>$; $c_2=<cond_2, b, \{F_2, F_3\}, 0>$; $c_3=<cond_3, c, \{F_1, F_2, F_3\}, 0>$;

$IP(t_i)= IP(t_j) = IP(t_k)=\{p_i\}$, $OP(t_i)= OP(t_j) = OP(t_k)=\{p_j\}$; $p_i=<\{a, b, c\}, 1>$, $p_j=<\{d, e, f\}, 1>$;

$F_1=<p_i, Pro_1>$, $F_2=<p_i, Pro_2>$, $F_3=<p_i, Pro_3>$, $F_4=<p_j, Pro_4>$, $F_5=<p_j, Pro_5>$, $F_6=<p_j, Pro_6>$;

$D(t_i)=<D_{min}(t_i), D_{max}(t_i), S(t_i)>$, $D(t_j)=<D_{min}(t_j), D_{max}(t_j), S(t_j)>$, $D(t_k)=<D_{min}(t_k), D_{max}(t_k), S(t_k)>$。

The calculate method of composition scheme generating probability is shown in Table 1.

4 Simulation Result

In this section, we introduce an application example to illustrate the application of CP-TCPN. As shown in Fig.4, in Stage 1, the user sends a request "*a*" from the terminal, and then the system will process the user's request so that the request will be transformed into a proper data series, which is named "*b*". At the same time, some related components will judge the user's terminal type: "*c*" or "*d*". Then different process method will be given. In Stage 2, the system will generate the users' location information and gather the services that may be needed in the future. The location information "*f*" is gained by the positioning system according to the request "*e*". After the request and location information is collected, the system will generate the composition scheme "*M&N"* or "*G&H"* according to the sorting method "*m&n*" or "*g&h*" (such as price and distance). These schedule schemes "*I&J*" or "*i&j*" are achieved by connecting the location information with the generated composition scheme.

In this example, we suppose the time of each transition's fire duration is 0.05 minimum and 0.15 maximum. According to Eq. (1), the firing duration's probability-distribution function of each transition can be achieved. The concrete method of the strong schedulability analysis and the probability of schedule scheme generation of the model described above can be further described as follows.

A. Strong schedulability analysis

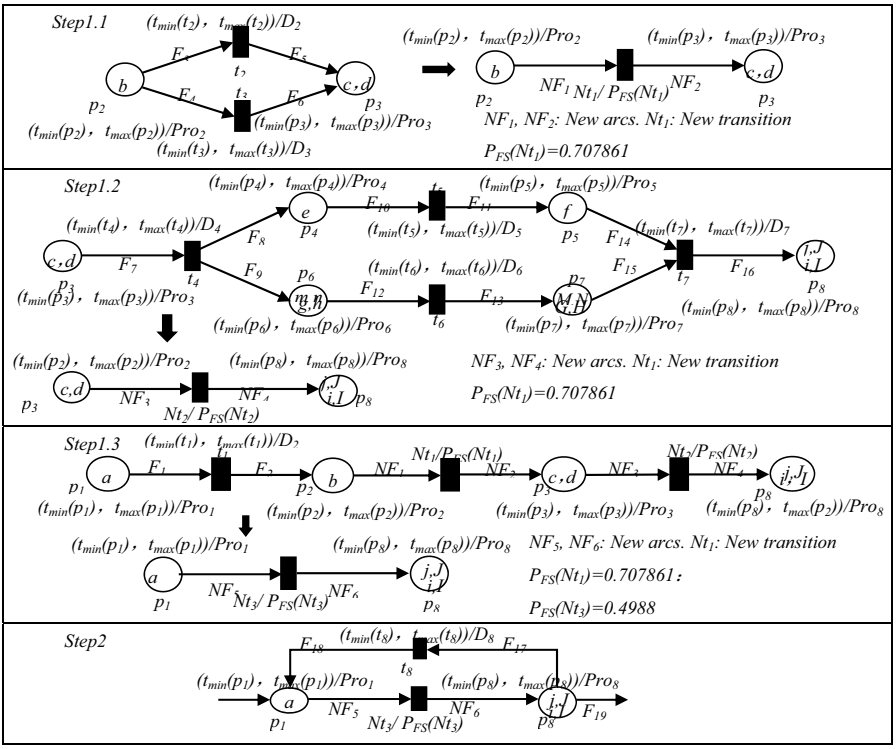
The probability calculation of every transition's strong schedulability is computed according to Eq. (9). We can linearize the nonlinear model structure and calculate the probability of the new model's reachability through the model's simplification as shown in Table 2, and the reachability probability of the whole model $P_{FS}(M)$ is as following:

$$P_{FS}(M)=P_{FS}(Nt_3)\times P_{FS}(t_8)=0.498279$$

B. The probability computing method for the schedule scheme's generation

During the process of services generation, there are four schedule schemes $\{I, i, J, j\}$. The generating relation of each token is shown in Fig.4:

Table 2. Model's simplification



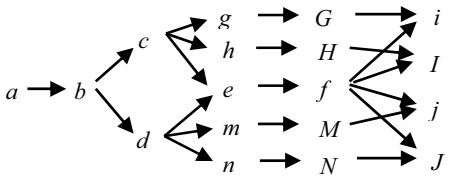


Fig. 4. The token generating relation

Table 3. The probability of each model's parameters

Condition	*cond*	*cond*$_2$		*cond*$_3$		*cond*$_4$		*cond*$_7$		*cond*$_8$		*cond*$_9$	*cond*$_{10}$	*cond*$_{11}$	*cond*$_{12}$
(cond)	*P(cond)*	*1*		*1*		*1*		*1*		*1*		*1/4*	*3/4*	*3/5*	*2/5*
Color(C)	*C*	*b*		*c*		*d*		*g*	*h*	*m*	*n*	*G*	*H*	*M*	*N*
	P(c_i)	*1*		*1*		*1*		*1*		*1*	*1*	*1*	*1*	*1*	*1*
Arc(F)	*F*	F_3	F_4	F_7		F_7		F_{12}		F_{12}		F_{15}	F_{15}	F_{15}	F_{15}
	Output	*c*	*d*	*e*	*g&h*	*e*	*m&n*	*G*	*H*	*M*	*N*	Δ_1	Δ_2	Δ_3	Δ_4
	P(F_i)	*1/2*	*1/2*	*1*		*1*		*1*		*1*		*1*	*1*	*1*	*1*

In the computing process of the model schedule scheme's generation probability, in order to simplify the analysis process, we don't show the probability of all the arcs and colors in Table 3, because the key reason for the final result is just the choice of some arcs and tokens with specific color.

The simulation result indicates that we can achieve that the selection probability of Scheme I is the largest as shown in Table 4. Therefore, we can conclude that Scheme I is the most widely accepted. When we recommend services composition scheme, Scheme I is the first choice. As the factors of the model will be changing with dynamic adjustment, the generating probability of each schedule scheme needs to be updated constantly. Therefore, the result can reflect the quality and generate situation of each services composition or schedule scheme in real time.

Table 4. The generating probability of each schedule scheme

Output token(p_j)	
token	*probability*
{i}	$P(i)= 1^4\times[1^2\times(1/2)\times 1^5\times(1/4)\times 1^2]=1/8$
{I}	$P(I)= 1^4\times[1^2\times(1/2)\times 1^5\times(3/4)\times 1^2]=3/8$
{j}	$P(j)= 1^4\times[1^2\times(1/2)\times 1^5\times(3/5)\times 1^2]=3/10$
{J}	$P(J)= 1^4\times[1^2\times(1/2)\times 1^5\times(2/5)\times 1^2]=2/10$

5 Conclusion

By using the tokens with specific colors in this study, we have proposed CP-TCPN based on TCPN by introducing probability. After expounding the differences between CP-TCPN and the traditional TCPN, we redefine a set of related parameters, such as the place and arc. The probability computing method of the model's schemed generation and strong schedulability modeled by CP-TCPN has also been discussed in detail. Finally, an application example with simulation result has been presented to show the usage of CP-TCPN in real situations. As it is shown in this paper, CP-TCPN can quantitatively analyze the model's schedulability through probability factor and dynamic composition scheme generations.

Acknowledgment. This work is partly supported by Shanghai Leading Academic Discipline Project (Project Number: J50103), and Natural Science Foundation of Ningxia under Grant No. NZ12212.

References

[1] Cardoso, J., Valette, R., et al.: Possibilistic Petri Nets. IEEE Transaction on Systems, Man, and Cybernetics-Part B 29(5), 573–581 (1999)
[2] Jin, Q., Vidale, R.F., Sugasawa, Y.: Optimum Order Time for a Spare Part Inventory System Modeled by a Non-Regenerative Stochastic Petri Net. IEICE Trans. Fundamentals E83-A(5), 818–827 (2000)
[3] Tsai, J.J.P., Yang, S.J.: Timing Constraint Petri Nets and Their Application to Schedulability Analysis of Real-Time System Specifications. IEEE Transaction on Software Engineering 21(1), 32–49 (1995)
[4] Feng, L., Jiang, H.: Analysis of Workflow Schedulability Based on Time Constraint Petri Nets. In: Proc. 10th International Conference on Computer Supported Cooperative Work in Design, pp. 1–5 (2006)
[5] Yu, R., Huang, Z., et al.: Modeling and Analyzing Project Performance with Timing Constraint Petri Net. In: Proc. ICCET 2009, vol. 2, pp. 243–246 (2009)
[6] Tsai, J.J.-P., Yang, S.J.H., Chang, Y.-H.: Schedulability analysis of real-time systems using timing constraint Petri nets. In: Proc. CDMP 1993, pp. 375–382 (1993)
[7] Hao, J., Sun, Z.-J.: The TCPN-Based Verification of Temporal Consistency in Web Service Process. In: Proc. ICEBE 2006 (IEEE International Conference on e-Business Engineering), pp. 302–306 (2006)
[8] Tsai, J.J.-P., Yang, S.J., Chang, Y.-H., Juan, E.Y.T.: Verifying timing properties for distributed real-time systems using timing constraint Petri nets. In: Proc. COMPSAC 1996, pp. 36–40 (1996)

An Implementation of Augmented Reality and Location Awareness Services in Mobile Devices

Pei-Jung Lin[1], Sheng-Chang Chen[2], Yi-Hsung Li[3], Meng-Syue Wu[1], and Shih-Yue Chen[1]

[1] Department of Computer Science and Information Engineering, Hungkuang University, Taiwan
[2] Ursa Pictor Design Co., Taiwan
[3] Institute for Information Industry, Taiwan

Abstract. The popularization of smartphones and advances in location-based technology have led to the creation of many applications and diverse mobile cloud technologies. Augmented reality (AR), which integrates virtual reality with the real world, is one of the mobile service technologies that have been receiving considerable attention in recent years. This study focuses on AR technology, which in conjunction with point of interest (POI) information established in a cloud database, enables users to instantly obtain services with the camera lenses of their mobile devices. The developed system allows users to quickly share AR images and information with others in their social networks from their current locations. This work includes social communities, photographing, radar detection, and GPS positioning that facilitate various human-machine interactions and information searches.

Keywords: Location-based Services, Mobile Social Network, Augmented Reality, Clouding Computing.

1 Introduction

With the increasingly wide application of mobile devices, smartphones and tablet computers are rapidly becoming integrated into every aspect of life, encompassing commercial marketing, scientific and technological industries, car navigation, interactive education, and game development. The scope of application is considerably vast and offers many advantages. In recent years, a number of technologies associated with augmented reality (AR) have extended to mobile devices. This technology is incorporated real-world elements into virtual reality. In AR, computer text and image information superimposed on images of the real-world environment provide accurate sensory information when users require it. Users can roam through a given space using mobile devices, integrating interaction between the virtual and the physical. AR is eye-catching in that it emphasizes the interaction between objects and real space. The applications and commercial opportunities for this technology are considerable. In this study, we developed a cloud database system integrating AR technology and social networking. The purpose of this system is to provide users with a location-based

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_76,

service (LBS) platform that supplies employment information. When users aim the camera lenses of their mobile devices at their surroundings, photographs are sent to the server site using cloud computing. The server site then retrieves and sends back data from the database, which is presented on the screens of users' smartphones or tablet computers, enabling them to access point of interest (POI) information corresponding to their GPS locations. Subsequently, the screens display job openings within their vicinities, presenting the information in AR message boxes. This is an extremely convenient system for those seeking employment.

According to a smartphone market-forecasting report released by the international research advisory firm, Gartner, 1.83 billion users are expected to be using their smartphones for internet access by 2013. For the first time, exceed that of users accessing the internet by computer (1.78 billion). The report also predicted that by 2014, over 70 billion downloads of mobile applications from App stores will be conducted yearly. In the same year, 20% of sales agents will be communicating through social networking services instead of through email, and personal cloud services will replace personal computers, forming the core of the digital industry. By 2015, approximately half of all online sales will originate from social networks and smartphone applications, and the data storage demands of consumers will be satisfied by free social network sites. At present, users are already accustomed to uploading a large number of photographs to Facebook. Smartphone usage is growing exponentially, of which Android and iOS hold a total market share of over 68%. With the development potential of this significant trend in mind, we developed a cross-platform mobile employment system using AR technology. This system can be installed in mobile devices using the two major operating systems mentioned above. We also integrated the social network Facebook, enabling users to share screen information. The popularization and convenience of mobile devices and the integration of AR [5, 8] in this study enable users to effectively satisfy their job-hunting needs. During the process of job-hunting, people generally utilize the internet to seek for job vacancies, by searching employment websites for example. However, it is hard to provide the job-hunting service according to users' location. We hope that the vast community of mobile device users can benefit from the convenience of our proposed mobile job-hunting system integrated with AR, and have access to employment information at any time.

2 Related Works

AR was developed from virtual reality technology, the concept of which originates from the "ultimate display" proposed by Professor Ivan E. Sutherland in the 1960s. In the following decades, virtual reality has been extended in the form of computer graphics, computer simulation, artificial intelligence, and sensing technology. Virtual reality uses virtual three-dimensional (3D) images generated by computer graphics to create a virtual environment that integrates the visual, auditory, and tactile sensory information. Although derived from physical reality, virtual reality does not allow users to perceive their physical surroundings. AR was thus created. It enables the virtual and the physical to co-exist in the same space. AR does not replace the real world; it augments it with virtual images and allows users to perceive the virtual world and the real world at the same time. AR [4, 6] comprises computer-generated virtual

images that use real-world objects as location coordinates. The images overlap or are merged to create a mediated reality that resides between the virtual and the real. Using the newly created mediated images, images of the real world can be augmented (or modified or diminished). Users can obtain corresponding sensory information [4, 9] through related devices (head-mounted displays, retina displays, or smartphones).

There are currently a number of existing platform systems for AR. Junaio [12], for example, displays AR images in the form of message boxes and uses radar detectors to conduct radar searches. This system is primarily used to plan transportation routes. Smartphone pictures and videos can also be stored in databases or uploaded and shared with social communities. Layar [11] enables software engineers to create various functions; users can see AR images with automatic message box displays, make phone calls, send e-mail, and plan routes. The Sekai Camera system [14] enables users to design their own AR; the customized functions allow users to input POI information to their current locations in the form of text or self-designed images. Other systems are designed for specific subjects. Wikitude [13], for instance, is designed for backpackers and enables users to search for recreation information and geological information on Wikipedia. Libre Geo Social [15] locates landmarks near the location of the user and provides select AR information; if the user selects fast food restaurants, the system will immediately show nearby fast food restaurants. Comparing the above, we found Junaio to be the most suitable application for job-hunting, as it allows users to limit the range of their search by distance. When in map mode, Junaio displays the locations of companies. Keywords can also be used to search for other subjects. Location-based job-hunting is thus enabled, and through the AR images, users can contact companies in which they are interested via phone or e-mail.

3 System Overview

Figure 1 displays the four major functions of the system: (1) AR Presentation: the opening screen of the system shows an introduction to the system with a reality virtual image, (2) Tagged Info: users can detect their own locations on maps, and in conjunction with multi-point positioning functions, they can plan routes on foot, by car, or by public transportation, (3) Social Network Sharing: using their smartphones, users can upload their photographs to social network sites and share them with friends, and (4) Radar Detection: a radar map displays nearby POI markers and message boxes containing distance-related information. These are the four core functions of the proposed system, which supports two major platforms, Android and iOS.

The developed system uses GPS to provide location-based information. In this case, the information involves job vacancies in the near vicinity. When users aim the lenses of their smartphone camera at the POI, virtual message boxes are displayed containing the marker information on companies or stores sent from the cloud database to the server site. POIs are locations of interest marked by users with appropriate information corresponding to the coordinates.

The information may include advertisements, URLs, or company contact details. The server site then returns the information to the mobile devices through Internet, enabling users to perform GPS tracking [1]. Job vacancies are displayed in AR

message boxes based on the location of the users, thereby fully utilizing the characteristics of virtual images and integrating them with employment information. The POI information in this study was obtained from employment websites, providing users with up-to-date information. Furthermore, mobile device users can download the system from Android Market or App stores.

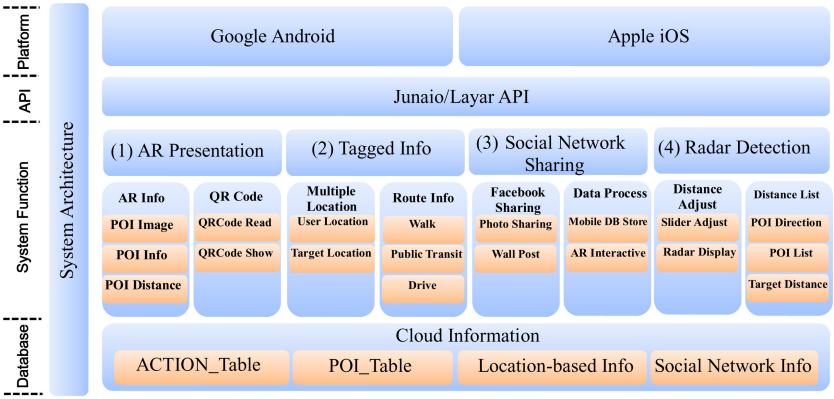
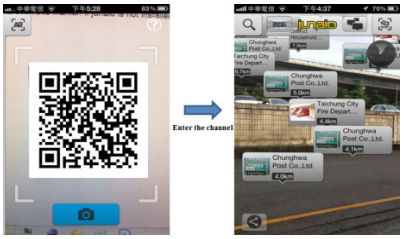


Fig. 1. Overview of system functions

3.1 AR Images

In the augmented virtual images on a smartphone, two relevant functions can be seen, as shown in Figure 2. The screen shows the virtual image provided by AR; the image presented by GPS positioning and the lenses merges the virtual with reality, making the message boxes more real [10]. Users can enter the developed employment system by simply scanning a QR Code with the mobile device, as seen in Figure 2 (a). Upon entering AR mode, the content of each message box is displayed using AR virtual images. When users select a message box, an information window will pop out, containing a brief introduction for users to browse. Other functions such as making a phone call, sending e-mail, playing videos, and planning routes are also shown. Users can also link to the URLs of company websites for further information, and share the information or screenshot with friends so that while integrating the information that user's need [2], the AR also enhances interaction with other users.

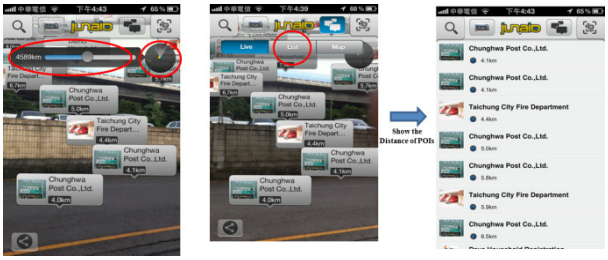


(a) (b)

Fig. 2. (a) QR Code scanning; (b) Actual screenshot

3.2 Radar Detection

This function displays all of the POI information in the near vicinity of users on radar maps on the mobile devices with the locations of the users in the center. This allows users to see their relative positions with each POI marker in addition to the pertaining information. The smartphone screenshots clearly indicate the markers nearest the users. As shown in the upper right hand corner of the screen in Figure 3(a), the radar map displays the POI markers as dots. Densely distributed dots show where POIs are high in number.



(a) (b)

Fig. 3. (a) Distance adjustment function; (b) List mode

As shown in Figure 3(a), the distance in the radar map is calculated in kilometers. When users take photographs, they can set a distance limit by selecting the radar map. This displays a scroll which shows the number of kilometers and with which users can adjust the distance range. When there is a greater number of AR message boxes, users can narrow the range to reduce the number of overlapping message boxes on their screen, thereby giving them a clearer screen image. By adjusting the distance on the screen, the message boxes will also change accordingly to demonstrate the distance between the POI and the users, as will the number of message boxes.

Figure 3(b) shows the list mode that users can select. The list mode lists the markers nearest the users and their distance from the users. In the event that the AR message boxes are very dense, users can switch to this mode. As users move, the content of each marker will change accordingly [7]. Using the distance adjustment function, users can control the amount of information in the list. The system also has a gyrocompass function that provides further accuracy in navigation.

4 Conclusions

The applications of AR are becoming increasingly diverse. In particular, the use of AR virtual images in smart mobile devices is rapidly growing in popularity. Our developed system combines four major functions: augmented reality, map search, social networks, and radar detection. Route planning for multiple means of transportation is also provided. To satisfy the needs of individuals seeking employment, the system uses the current location of users to provide POI information with LBS. This allows users access to employment information anytime, anywhere. We tested the developed system on two major operating systems: Apple iOS and Android. The AR screenshots

can uploaded to social network sites and shared with friends instantly. As the number of individuals using mobile devices such as smartphones and tablet computers increase, an AR job-hunting application with LBS can enhance the efficiency of searching for this kind of information.

Acknowledgment. The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 101-2221-E-241-012.

References

1. Ajanki, A., Billinghurst, M., Gamper, H., Järvenpää, T., Kandemir, M., Kaski, S., Koskela, M., Kurimo, M., Laaksonen, J., Puolamäki, K., Ruokolainen, T., Tossavainen, T.: An Augmented Reality Interface to Contextual Information. Journal of Virtual Reality 15(2-3), 161–173 (2011)
2. FitzGerald, E., Adams, A., Ferguson, R., Gaved, M., Mor, Y., Rhodri, T.: Augmented Reality and Mobile Learning: The State of The Art. In: Proceedings of 11th World Conference on Mobile and Contextual Learning (mLearn 2012), Helsinki, Finland, October 16-18 (2012)
3. Jang, S.H., Andrew, H.S.: GIS and Augmented Reality in 2015. Association of Geographic Information (AGI) Foresight Study (February 5, 2010)
4. Kent, J.: The Augmented reality Handbook - Everything you need to know about Augmented reality. Emereo Pty. Ltd. (April 21, 2011)
5. Madden, L.: Professional Augmented Reality Browsers for Smartphones: Programming for junaio, Layar and Wikitude. Wrox Press (July 9, 2011)
6. Narzt, W., Pomberger, G., Ferscha, A., Dieter, K., Reiner, M., Jan, W., Horst, H., Christopher, L.: Pervasive Information Acquisition for Mobile AR-navigation Systems. In: Proceedings of 5th IEEE Workshop on Mobile Computing Systems & Applications (WMCSA 2003), Monterey, CA, USA, October 9-10, pp. 13–20 (2003)
7. Pomberger, G.: Digital Graffiti - A Framework for Implementing Location-Based Systems. International Journal of Software and Informatics 5(1-2), 355–377 (2011)
8. Schinke, T., Henze, N., Boll, S.: Visualization of Off-Screen Objects in Mobile Augmented Reality. In: Proceedings of 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI 2010), September 7-10, pp. 313–316 (2010)
9. Specht, M., Ternier, S., Greller, W.: Mobile Augmented Reality for Learning: A Case Study. Journal of the Research Center for Educational Technology 7(1) (2011)
10. Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T., Schmalstieg, D.: Real-Time Detection and Tracking for Augmented Reality on Mobile Phones. IEEE Transactions on Visualization and Computer Graphics 16(3), 355–368 (2010)
11. Layar, `http://www.layar.com/`
12. Junaio, `http://www.junaio.com/`
13. Wikitude, `http://www.wikitude.com/`
14. Sekai Camera Web, `http://sekaicamera.com/`
15. LibreGeoSocial: Augmented Reality FLOSS
16. `http://www.libregeosocial.org/`

Development of STEAM Education Program Centering on Non-traditional Energy*

Yilip Kim[1], Jeongyeun Kim[1], Namje Park[1,**], and Hyungkyu Lee[2]

[1] Major in Computer Education, Faculty of Science Education,
Graduate School, Jeju National University,
61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Korea
{yilipkim,namjepark}@jejunu.ac.kr, inarasam@naver.com
[2] Electronics and Telecommunications Research Institute (ETRI),
218 Gajeong-ro, Yuseong-gu, Daejeon, 305-700, Korea
leehk@etri.re.kr

Abstract. The purpose of this paper is as follows. First, it is to develop a STEAM education program centering on non-traditional energy (gas hydrate and shale gas) targeting high school students. Second, it is to develop teaching-learning materials that can be used in the education program. To achieve the purpose of this paper, we investigated the contents of the subject using newspaper articles, academic books, and academic journals and also analyzed the curriculum and textbooks of related subjects revised in Korea's 2009 to determine the learning capability of the students by age.

Keywords: STEAM, Non-traditional Energy, Elementary School, Teaching Method.

1 Introduction

For the reserves of currently used thermal energy sources such as petroleum and coal are limited, the gas hydrate and shale gas which are classified as non-traditional gases are recently drawing attention as the alternative energy source. These gases are found in the stratum, and a number of countries in the world are showing interest in the economic feasibility of these gases as the technology for extracting this gas has advanced recently. Korea also became interested in the use and extraction of the shale gas, and we decided to develop a STEAM(Science, Technology, Engineering, Art, Mathematics) program on the subject of the shale gas targeting high school students based on the idea that the integrated learning of various subjects in relation to energy and its use in the school curriculum can be efficient for both learners and teachers.

The purpose of this paper is as follows. First, it is to develop a STEAM education program centering on non-traditional energy (gas hydrate and shale gas) targeting

* This work was supported by the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy(MKE, Korea). [10038653, Development of Semantic based Open USN Service Platform].

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_77,

high school students. Second, it is to develop teaching-learning materials that can be used in the education program. To achieve the purpose of this paper, we investigated the contents of the subject using newspaper articles, academic books, and academic journals and also analyzed the curriculum and textbooks of related subjects revised in 2009 to determine the learning capability of the students by age.

2 Purpose of Program Development

We held a number of seminars among us and a consultative meeting with outside experts to review the feasibility of the contents and the applicability of the education program development plan we developed. Furthermore, the teaching materials to be applied to the education program were produced to reflect convergence in contents domain pursued by STEAM education and learner's characteristics in terms of educational psychology.

The context (Co), the framework of the convergent education (STEAM), dealt with the utilization status of fossil fuels, its future prospect, and issues related with the use of energy. Creative Design (Cd) made students in teams devise models for collecting gas hydrate and shale gas. Emotional Touch (ET) made students to touch various materials when developing a model for gas collecting device and compose them close to actual situation in order for them to experience various emotional touches.

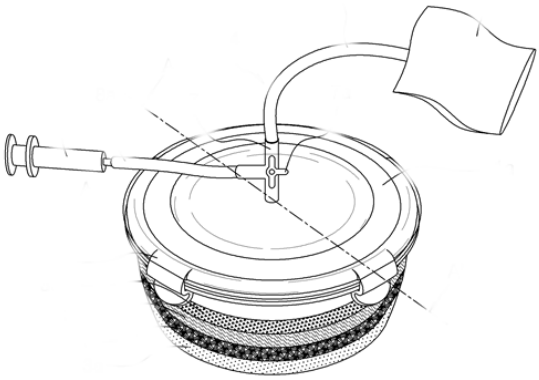


Fig. 1. Teaching Aid to be applied to STEAM Program and the Completed Model

Major outcomes of this paper are as follows. First, the non-traditional energy theme was composed of 10 classes in total centering on gas hydrate and shale gas, and major contents of each class were the following. Understanding non-traditional energy (1st class), Understanding gas hydrate: Problem solving (2nd class), Experiment of the characteristics of gas hydrate using alternative materials (3rd class), Design of gas collector for gas hydrate (4th class), Development and evaluation of gas hydrate collector (5th class), Understanding shale gas (6th class), Experiment of the characteristics of shale gas using alternative materials (7th class), Designing the stratum containing shale gas and the collector (8th class 8), Making the model for

the stratum containing shale gas and the collector (9^{th} class), and Production of gas using the model for the stratum containing shale gas and the collector (10^{th} class).

Second, we developed a teaching aid, the teaching and learning materials that will be applied to the program to allow students to examine the characteristics of shale gas contained in the stratum and directly extract the gas. This teaching aid allows students to check the stratum that contains the shale gas which is one of the non-traditional energy sources and apply various strategies to extract the shale gas and evaluate the result through problem solving exercises using alternative materials (i.e. dry ice or soda water).

Third, teaching-learning course guide for teacher which will provide the guideline for teaching and learning during the implementation of 10 classes was developed along with the instruction guide for teacher and textbooks for students.

3 Method of STEAM Program Development

To achieve the objectives of the paper, we investigated the subject using newspaper article, academic books, and academic papers and also analyzed curriculums and textbooks revised in 2009 in related subjects in order to determine the learning capability of students by age.

We held a number of seminars among us and a consultative meeting with outside experts to review the feasibility of the contents and the applicability of the education program development plan we developed. Furthermore, the teaching materials to be applied to the education program were produced to reflect convergence in contents domain pursued by STEAM education and learner's characteristics in terms of educational psychology.

The context (Co), the framework of the convergent education (STEAM), dealt with the utilization status of fossil fuels, its future prospect, and issues related with the use of energy. Creative Design (Cd) made students in teams devise models for collecting gas hydrate and shale gas. Emotional Touch (ET) made students to touch various materials when developing a model for gas collecting device and compose them close to actual situation to experience various emotional touches.

4 Result of STEAM Program Development

To apply on site without creating conflicts with current curriculum, a systematic connection into the basic science technology engineering is required. In addition, integrative thought

4.1 Teaching Aid for Class

- Teaching Aid for Class: A model for learning about the stratum containing shale gas and the extraction system
- Characteristics of Teaching Aid: Students can see the stratum that contains the shale gas which is one of non-traditional energy and apply various strategies

concerning the extraction of shale gas while solving problems about alternative materials (i.e. dry ice or soda water) and also evaluate the result.

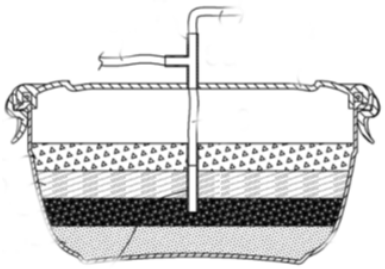


Fig. 2. A model for learning about the stratum containing shale gas and the extraction system

5 Proposed STEAM Program

Grade/Classes		Grade 10 / 10			
Implementation Model		Theme-oriented learning, problem-based learning, project learning			
Class	Stage	Sub Theme	Major Contents		Related Subject
1	Check Theme	What is non-traditional energy?	Learning Activities	Activity 1: Learning about non-traditional energy Activity 2: Sharing opinions about the need of non-traditional energy	Science
2	Investigation and Research Activities	Let's learn about gas hydrate.		Activity 1: Introduction to the characteristics of gas hydrate Activity 2: Introduction to the extraction of gas hydrate Activity 3: Design of extraction scenario per team (Theme: I would extract it this way.)	Science, Technology
3		Design gas hydrate extractor using a substitute.		Activity 1: Presentation of extraction scenario designed last time Activity 2: Suggest test materials under similar environment as gas hydrate and design test	Technology, Engineering, Arts

4 ~ 5		Produce a gas hydrate extractor model using a substitute and test.		Activity 1: [Warm Water vs. Cold Water] Which method can collect more gas? Activity 2: [Salt Water vs. Fresh Water] Which method can collect more gas?	Science, Technology
6		Let's learn about shale gas.		Activity 1: Introduction to the characteristics of shale gas Activity 2: Think about and discuss social issues about shale gas	Science, Technology
7		What are the technological developments related with shale gas?		Activity 1: Introduction to shale gas extraction technology Activity 2: Guide to experiments using alternative materials Activity 3: Understanding the characteristics of the materials used in the experiment and consideration about the use of materials	Science, Technology
8~10		Produce a shale gas model using a substitute and test.		Activity 1: Basic setting for experiment Activity 2: Basic setting + design the best way to extract gas Activity 3: Conduct actual experiment based on the design (each team describes the method and conduct the experiment while other teams watch) and prepare for the presentation based on the outcomes Activity 4: Presentation about the advantage and disadvantage of the experiment and discussion	Science, Technology, Mathematics

6 Conclusion

This paper develops a STEAM education program centering on non-traditional energy (gas hydrate and shale gas) targeting high school students. To achieve the purpose of this paper, we investigated the contents of the subject using newspaper articles, academic books, and academic journals and also analyzed the curriculum and textbooks of related subjects revised in 2009 to determine the learning capability of the students by age.

References

1. Park, N., Ko, Y.: Computer Education's Teaching-Learning Methods Using Educational Programming Language Based on STEAM Education. In: Park, J.J., Zomaya, A., Yeo, S.-S., Sahni, S. (eds.) NPC 2012. LNCS, vol. 7513, pp. 320–327. Springer, Heidelberg (2012)
2. Ko, Y., An, J., Park, N.: Development of Computer, Math, Art Convergence Education Lesson Plans Based on Smart Grid Technology. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech, CA, CES3 2012. CCIS, vol. 339, pp. 109–114. Springer, Heidelberg (2012)
3. Ko, Y., Park, N.: Experiment and Verification of Teaching Fractal Geometry Concepts Using a Logo-Based Framework for Elementary School Children. In: Kim, T.-H., Adeli, H., Slezak, D., Sandnes, F.E., Song, X., Chung, K.-I., Arnett, K.P. (eds.) FGIT 2011. LNCS, vol. 7105, pp. 257–267. Springer, Heidelberg (2011)
4. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
5. Park, N.: Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
6. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
7. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)
8. Park, N., Kim, Y.: Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 193–202. Springer, Heidelberg (2010)
9. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 142–152. Springer, Heidelberg (2010)
10. Park, N.: Customized Healthcare Infrastructure Using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
11. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
12. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
13. Park, N.: The Implementation of Open Embedded S/W Platform for Secure Mobile RFID Reader. The Journal of Korea Information and Communications Society 35(5), 785–793 (2010)
14. Park, N., Cho, S., Kim, B.-D., Lee, B., Won, D.: Security Enhancement of User Authentication Scheme Using IVEF in Vessel Traffic Service System. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 699–705. Springer, Heidelberg (2012)
15. Park, N., Cho, S., Kim, B.-D., Lee, B., Won, D.: Security Enhancement of User Authentication Scheme Using IVEF in Vessel Traffic Service System. In: Yeo, S.-S., Pan, Y., Lee, Y.S., Chang, H.B. (eds.) Computer Science and its Applications. LNEE, vol. 203, pp. 699–705. Springer, Heidelberg (2012)

Scalable Key Management for Dynamic Group in Multi-cast Communication*

Fikadu B. Degefa and Dongho Won**

College of Information and Communication Engineering, Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon-si, Gyeonggi-do, 440-746, Korea
befikadub@yahoo.com, dhwon@security.re.kr

Abstract. To have secure multicast group communication, group key management plays an essential role to guarantee data security. Because communication bandwidth, storage memory, and computational power are limited resources, most group key management schemes for scalable secure multicast communications have focused on reducing the number of update messages, number of stored keys, and computational load. Here, also we propose efficient scheme in such a way that solves these problems.

Keywords: Key management, Dynamic group, Multi-cast.

1 Introduction

Secure and reliable communications have become critical in modern communication system. Secure distribution of copyright-protected material (e.g. music), and audio streaming are all methods of multicasting to a large number of users.

In many commercial applications of multicast and broadcast, it is desirable that only those users who have paid for the service can retrieve broadcast data. Thus, it needs basic cryptographic requirements [1] such as: data confidentiality, data integrity, authentication, and access control are required to build secure collaborative system in the multicast channel. Encryption solves the problem of confidentiality and secures the group communication, but it creates the problem of managing the encryption/decryption keys to the group members securely and efficiently. Assuming the use of strong security mechanisms that cannot be easily defeated by frivolous cryptanalytic attacks, we focus our security concerns and the rest of our technical discussion around key material management mainly about key distribution for dynamic multicast groups.

The rest of our paper is organized as follows: section II comes up with newly proposed scheme for dynamic multicast group Security. Security and Performance analysis will appear in section III and IV respectively .The finally section is conclusion.

* This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.

** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_78,

2 Proposed System

2.1 Motivation of the Proposed Scheme

- Minimizes hardware and software implementation complexity
- It needs comparatively less communication and computational cost.
- Reducing overall delay of key generation and distribution by reducing computational load on group controller
- It enables member join and member leave very simple.

2.2 System Requirements

Dynamic multicast groups are distributed and inherently dynamic in nature in which multiple group members may join, leave, and be evicted simultaneously .Thus, group key management is often complicated due to high group dynamics. In order to have secure and efficient scheme for dynamic multicast groups' key distribution, we considered Scalability [5], Reliability [7], Forward Secrecy [2][6], Backward Secrecy[2][6], Collusion Resistance[6] and Power Proximity[5] as basic requirements.

2.3 Assumptions

We assume that a complete m-ary tree is a rooted tree in which all leaves have the same depth and all internal nodes have m children. The protocol requires two system parameters: n and g. Both parameters are considered as public values and may be used by all users in the system. Parameter n is a prime number (usually very large) and parameter g (also called a generator) is an integer smaller than n with an assumption of one way function application .A reader should also keep in mind that all subgroup members are trusted.

2.4 System Model

2.4.1 Initialization

One of the goals of this protocol is to reduce the overload of the group controller; hence, we design our scheme in such a way that shares computational and communication load to subgroup controllers. As a central idea, we divide the multicast communication group into local subgroups, which is independently managed by a subgroup controller (SGC) like a separate multicast group. Thus, when a member joins or leaves the communication group, it joins or leaves only its local subgroup. As a result, only the local subgroup communication key needs to be refreshed and the scalability problem is greatly mitigated. We consider that every node is a subgroup controller to its children and a subgroup member of its parent node. The remaining steps of the key generation are briefly discussed as follows:

Step1*:*
Initially SGCs(including KGC) form a two-party group with each of the remaining group members using Diffie-Hellman technique. First, KGC selects a private key x_c

and generates a public Key X_c($X_c = g^{X_c} \bmod n$) and broadcasts to the subgroup members($SGC_{2,1}$ to $SGC_{2,m}$). Each subgroup member $SGC_{2,j}$ also assumes a private key and generates a public key as $X_{2,j} = g^{x_{2,j}} \bmod n$, Where $x_{2,j}$ is the private key of $SGC_{2,j}$ for $1 \le j \le m$. Each $SGC_{2,j}$ then transmits $X_{2,j}$ to KGC and its respective children ($SGC_{3,j}$s). After exchanging the public keys, each member generates a unique shared key with KGC as $K_{2,j} = (X_c)^{x_{2,j}} \bmod n = g^{x_c x_{2,j}} \bmod n$ (for $i \le j \le m$). Similarly, KGC generates the same shared key using $K_{2,j} = \left(X_{2,j}\right)^{x_c} = g^{x_c x_{2,j}} \bmod n$, (for $1 \le j \le m$). The KGC actually generates (m) shared key $K_{2,j}$s for (m) parties for $1 \le j \le m$.

Subsequently, every $SGC_{2,1}$ to $SGC_{2,m}$ generates two-party key with their respective subgroup members (i.e. $SGC_{3,1}$ to $SGC_{3,m}{}^{2)}$ by the same procedure as KGC. i.e. Using the same fashion — every subgroup controller $SGC_{i,j}$ _ forms two-party key with its members, for $1 \le i \le L\text{-}1$ and $1 \le j \le m^{L-2}$.

Step 2:

After the two-party key is generated within each controller and subgroup, the smallest two-part key of all the subgroups is identified by bottom- up searching algorithm.

i. Initially, every leave node ($SGC_{L-1,1}$ to $SGC_{L-1,m}{}^{L-2}$) compares its subgroup members' two-party keys($K_{L,1}$ to $K_{L,m}{}^{L-1}$) and returns the smallest to its parent SGC .
ii. The parent SGC compares the returned results from its children and again returns the smallest to its Parent SGC.
iii. Following the same procedure, the smallest of all subgroups two-party key is identified.
iv. Eventually, the KGC selects a private number smaller than the smallest two-party key as a group data encryption key(DEK).

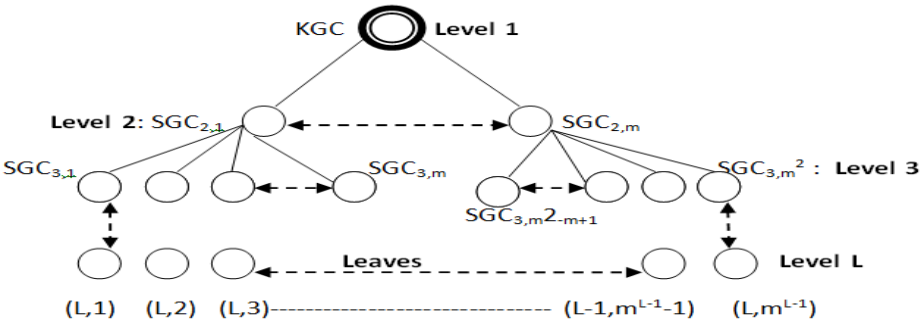


Fig. 1. Subgroup controller based key distribution tree

***Step 3*:**

The KGC performs a multiplication and addition operation to share the group data communication key (DEK) for each child node (SCG) using formed two-party keys. KGC computes Z as $Z = K_{2,1} * K_{2,2} \ldots\ldots * K_{2,m} + DEK = (\Pi K_{2,j}) + DEK$ (for $1 \le j \le m$) and Z is broadcasted to all KGC's children that are $SGC_{2,j}$ s, for $1 \le j \le m$.After receiving Z, each child node $SGC_{2,j}$ computes DEK= (Z mod $K_{2,j}$, ,(for $1 \le j \le m$). Again each

$SGC_{2,j}$ computes $Z_{2,j}$ as the same fashion as KGC do, using its own members two-party unique keys(for$1 \leq j \leq m$) and then broadcasts to $SGC_{3,j}$s,

1) $Z_{2,1'}=K_{3,1}*K_{3,2}\ldots\ldots*K_{3,m}+DEK = (\Pi K_{3,j})+DEK$,(for$1 \leq j \leq m$).
2) $Z_{2,2}=K_{3,m+1}*K_{3,m+2}\ldots\ldots*K_{3,2m}+DEK= (\Pi K_{3,i})+DEK$,(for $m+1 \leq j \leq 2m$)

……

m) $Z_{2,m}=K_{(3,m^2-m+1)}*K_{(3,m^2-m+2)}\ldots\ldots*K_{(3,m^2)}+DEK = (\Pi K_{3,j})+DEK$,(for $m^2-m+1 \leq j \leq m^2$).

$SGC_{3,j}$s recovery the data encryption key (DEK) as follows:

- SGC3,1 computes DEK= $Z_{2,1}$ mod $K_{3,1}$

- $SGC_{3,m}$ computes DEK= $Z_{2,1}$ mod $K_{3,m}$
- $SGC_{3,m+1}$ computes DEK= $Z_{2,2}$ mod $K_{3,m+1}$

- SGC_{3,m^2} computes DEK= $Z_{2,m}$ mod K_{3,m^2}

Accordingly —every communication member— receives the group shared key DEK.

2.4.2 Member Join

A member join is an easy job in which DEK with the controller private key are the only changed keys. Let a member requests to join the communication, is inserted in a position of Subgroup controller as in Fig2 New_1, the controller to which the new member belongs to ($SGC_{2,2}$) and the new member form a two-part key. Also the newly join subgroup controller creates m two-party keys with all of its subgroup members. Then after, the smallest two-party key of the newly created two- party keys is returned to the KGC for comparison through two- party (because using the old group key to return the value, compromises the backward secrecy) based communication.

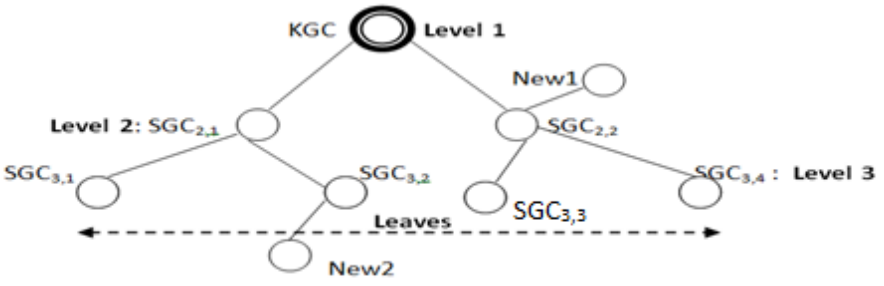


Fig. 2. Sample of the new subgroup controller based tree scheme

To visualize this concept, let us add the following expressions:

- The new $SGC_{2,2}$ selects private key $x_{2,2}$ to generate public key $X_{2,2.}$
- $X_{2,2} = g^{x_{2.2}} \ mod \ n$

The new $SGC_{2,2}$ broadcasts $X_{2,2}$ to KGC and to its children $SGC_{3,3}$ &$SGC_{3,4}$. It also acquires their public key from their public directory. This member forms the following two-party keys with KGC as well as with its children($SGC_{3,3}$ &$SGC_{3,4}$) and returns the smallest value of these keys ($K_{2,2}$, $K_{3,3}$ $K_{3,4}$) to KGC.

- $K_{2,2}=g^{x_c x_{2.2}} \ mod \ n$, $K_{3,3} = g^{x_{3,3} x_{2.2}} \ mod \ n$, $K_{3,4} = g^{x_{3,4} x_{2.2}} \ mod \ n$

But, if the insertion point is as in Fig. 2 New_2, the new member only creates one two-party key, which is returned to the KGC, with its parent node (controller). Finally, KGC selects new shared group key less than the old DEK and returned value.

2.4.3 Leave Member

If a leave member is a subgroup controller as in Fig. 3 $leave_1$,

- The new $SGC_{2,2}$ selects private key $x_{2,2\text{-new}}$(private key) from a prime number pool . Computing public key $X_{2,2\text{-new}}$, $SGC_{2,2}$ broadcasts the public key to KGC and to its children $SGC_{3,3}$ &$SGC_{3,4}$.
- $X_{2,2\text{-new}} = g^{x_{2,2-new}} \ mod \ n$
- Selecting a new private key $x_{c\text{-new}}$, KGC generates public key $X_{c\text{-new}} = g^{x_c - new} \ mod \ n$, then broad cast to $SGC_{2,1}$ and $SGC_{2,2}$ so as to form the following new two-party keys(Notice that we are looking $SGC_{2,2}$ as a specific example of a member leave).
- $K_{2,2\text{-new}}=g^{x_{c-new} x_{2,2-new}} \ mod \ n$, $K_{2,1\text{-new}} = g^{x_{c-new} x_{2,1}} \ mod \ n$
- $SGC_{2,2}$ also generates two-party key with its children($SGC_{3,3}$ &$SGC_{3,4)}$ and KGC
- $K_{2,2\text{-new}} = g^{x_{c-new} x_{2,2-new}} \ mod \ n$, $K_{3,3\text{-new}}=g^{x_{3,3} x_{2,2-new}} \ mod \ n$, $K_{3,4\text{-new}} =g^{x_{3,4} x_{2,2-new}} \ mod \ n$
- Subsequently, the smallest value of $X_{2,1\text{-new}}$, $K_{2,2\text{-new}}$, $K_{3,3\text{-new}}$, and $K_{3,4\text{-new}}$ is returned to KGC.

But if the case is as Fig. 3 $leave_2$, the subgroup controller to which the leave member belongs to($SGC_{3,2}$), selects a new private key to form two-party key with each of its members(leaves). Then after, the smallest two-party key of the newly created two- party keys is returned to the KGC using old group key for comparison in which KGC selects new shared group key less than the old DEK and returned value.

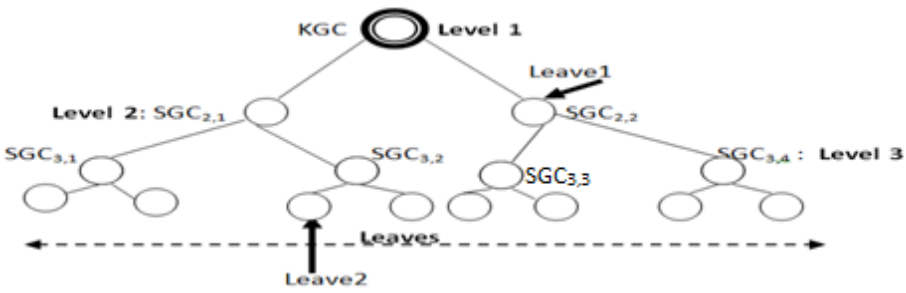


Fig. 3. Sample of the new subgroup controller based tree scheme (member leave)

As shown above, in every multi group subgroup based key management tree of our scheme, member leave is handled accordingly.

Finally, KGC redistributes the new shared group key in the same procedure as initialization phase.

2.4.4 Mass Leave

The behavior of the massive leave of this scheme is very similar to single member leave. The difference is that the number of private and two-party keys changed in the case of mass leave is proportional to the number of members leave the communication. As in the case of single leave, if the members to leave are subgroup controllers, new subgroup controllers are assigned from the members that have stayed in the communication for long period time and capable of managing according to the criteria set in advance.

3 Security Analysis

The security of our protocol is based on the assumption that each current group user is trusted that will keep group key secret, and each group controller will keep two-party key. Moreover, we assume that user authentication, source authentication key communication integrity and confidentiality are realized through asymmetric digital signature and certificate.

A fundamental challenge in designing secure hierarchical cryptographic key management schemes is protecting against attacks perpetrated by valid users in possession of correct key. In our protocol , we can conclude that a passive hacker who knows a contiguous subset of old or new group keys cannot discover any subsequent group key because there no any mathematical relation between the new and the old keys.

Moreover, when a member joins or leave, new key is created in such a way that it is hidden from the member. This ensures forward and backward secrecies in our scheme.

4 System Performance Comparison

This protocol focuses on the optimization of user computation, communication and number of the re-key message load. The summary of the performance analysis is as follows:

Table 1. Comparison Table

Key Management schemes	*Key update message*	*Storage overhead*	*Computational overhead (exponentials)*
Minimal key storage scheme	O(N)	O(1)	N/A
Logical key hierarchy scheme	O(logN)	O(N)	N/A
Hybrid tree structure scheme	O(M + log(N/M)	O(log(N/M))	N/A
Tree group Diffie-Hellman scheme	O(1)	O(log2N)	O(log2(N)
Our proposed scheme	O(1)	O(1)	O(log(N/M))

5 Conclusion

The key management scheme considers and takes an advantage of minimizing the computational overhead of inherently expensive cryptographic operations. Because of the complexity of group key generation, group key management needs to adopt a key tree structure in order to reduce group key generation times. Key trees have been suggested for centralized group key distribution systems in order to reduce the complexity of key calculation. Accordingly, we believe that our proposed scheme is reasonably applicable to many multicasting communication groups.

References

1. Kruus, P.S., Macker, J.P.: Techniques and issues in multicast security. In: Military Communications Conference 1998 Proceedings, pp. 1028–1032 (1998)
2. Srinivasan, R., et al.: Secure group key management scheme for multicast networks. International Journal of Network Security 11(1), 33–38 (2010)
3. Molva, R., Pannetrat, A.: Scalable Multicast Security with Dynamic Recipient Groups. ACM Transactions on Information and System Security 3(3), 136–160 (2000)
4. Naresh, V.S., Murthy, N.V.E.S.: Diffie-Hellman Technique Extended to Efficient and Simpler Group Key Distribution Protocol. International Journal of Computer Applications 4(11), 1–5 (2010)
5. Dong-Hyun, J.E., et al.: Computation-and-storage-efficient key tree management protocol for secure multicast communications. Computer Communications 33(2), 136–148 (2010)
6. Wallner, D.M., Harder, E.G., Agee, R.C.: Key Management for Multicast: Issues and Architecture, RFC 2627, IETF (1999)
7. Hardjono, T., Cain, B., Doraswamy, N.: A Framework for Group Key Management for Multicast Security. Internet Draft, IETF (1998)

Result of Implementing STEAM Program and Analysis of Effectiveness for Smart Grid's Education*

Jeongyeun Kim, Yilip Kim, and Namje Park**

Major in Computer Education, Faculty of Science Education,
Graduate School, Jeju National University,
61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Korea
{yilipkim,namjepark}@jejunu.ac.kr, inarasam@naver.com

Abstract. STEAM education allows students to improve themselves in cognitive and affective domains in math. science. Visualizing and storytelling science as described above is a good approach towards learning science in an easy and exciting way. Aware of such, our paper plans to advocate convergence curriculum to nurture students' creativity, problem-solving skills and ultimately support them to become a creative talent built on convergence. This paper intends to investigate the effect of STEAM education program using smart grid on elementary school students. For this, this paper implemented STEAM education program using smart grid for ten students in the 4th grade in Jeju Special Self-Governing Province.

Keywords: STEAM, Smart Grid, effectiveness, Science, Technology.

1 Introduction

STEAM is an acronym of Science, Technology, Engineering, Arts, and Mathematics. This is an educational curriculum that combined Art to the existing American STEM (Science, Technology, Engineering, Mathematics) curriculum and Yakman (2008) defined the STEAM education in the following two directions.

First, it is an education where Science, Technology, engineering, and mathetmatics include other areas in addition to the standards of their own and second, it is an integrative education that purposefully includes the actual subjects and teaching matters. For a more detailed definition, Yakman (2008) suggested the framework. As shown in the framework, the STEAM education determines the level from the lifelong learning to detailed academic content classification. The first level is the Lifelong learning. This stage signifies our adaptation to our surroundings and sustained learning that are unintended and unavoidable. Second level is Integrative learning. In this stage, the student learns the basic overview of all of the academic fields and how they are related. The best way is to learn by topics. This stage of

* This work was supported by the Korea Foundation for the Advancement of Science & Creativity(KOFAC) grant funded by the Korean Government.
** Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_79,

learning is appropriate for elementary and middle school education. The third level is Multidisciplinary learning. This stage allows the student to learn a specific field and how they are related to real life. The best method is to learn the practical uses. This is appropriate for middle school education.

The fourth level is Discipline learning. This stage focuses on each educational stage and is appropriate for middle school education. The last fifth level is Content Specific learning. This stage deals with detailed research of each field and is appropriate for high school and professional education.

Therefore, STEAM education does not entail a part of education but refers to an overall paradigm from the professional learning to lifelong learning, which is organized with the addition of art to the existing education, especially in the integrated education of Science, Technology, Engineering, Mathematics and Art in elementary school education. Our paper plans to advocate convergence curriculum to nurture students' creativity, problem-solving skills and ultimately support them to become a creative talent built on convergence.

2 STEAM Education Program Using Smart Grid

2.1 Approach of Research Method

The STEAM education program was implemented in 6 classes in a few weeks, and teachers were given instructions before the program was implemented to fully understand the contents of the program so that STEAM education is implemented in accordance with the objectives.

A questionnaire survey was conducted after the program ended to find out how effective the STEAM education program was to students by examining the level of difficulty, the level of students' interest in the program, and the relevance to school subjects.

1 class offered the overview of the theme, 3 classes were allocated for exploration and research, and 2 were allocated for processing the result of exploration and research to focus on the theme throughout the program. The program encouraged students to use real life cases and develop not only the knowledge but also the attitude, supplement insufficient activity hours through discussion after the activity, and design their own green home to realize self-directed learning.

Students were allowed to use various evaluation methods such as self-evaluation, peer-evaluation, portfolio, and observation-based evaluation.

2.2 STEAM Learning Materials Using Smart Grid

To implement STEAM education program on IT theme, particularly smart grid, for participation of elementary school students, a textbook for students were prepared. The textbook was prepared especially with a variety of stories, cartoons, pictures, and photos laid out in a storytelling format to draw students' attention and it was designed to promote self-directed learning. Also, the guide for teachers was developed for teachers to understand the contents to teach and for efficient teaching.

The textbook was revised and complemented through expert advice and has been continuously supplemented through the verification by experts since when it was first developed.

3 Proposed STEAM Education Program

This is example of STEAM convergence education. Contents are as follows.

- Theme: property of figures on axisymmetric location
- Goal: identify property of figures on an axisymmetric location and draw them

1) Activity 1. Learning about property of figures on an axisymmetric location
Guide students to draw figures like heptagon and decagon that cannot be easily drawn with a protractor by giving easy LOGO commands like 'forward' and 'right' by using Microsoft Logo Program.

2) Activity 2. Drawing congruent figures with LOGO
Check figures in the activity sheet and draw congruent figures to apply congruence principle learned from PC.

3) Activity 3. Drawing figures on axisymmetric location
Guide students to make a presentation on how to draw figures on an axisymmetric location. Paint various figures, fold them in half and check their Decalcomania form. Make figures on axisymmetric locations on a drawing paper, paint them and fold them in half.

In doing so, teachers should aid students to determine an angle that fits a specific figure when they draw figures by using Microsoft Logo Program rather than simply give the answers. Since drawing congruence of figures may be a challenge to students, teachers should walk around the classroom to help students. Also, they should encourage students to try by themselves, learn from it and digest the principles than merely teach them when introducing the principle of Decalcomania.

- Theme: apply the concept of smart grid by making a simple smart grid
- Content: consumption of electricity and power is about electric consumption while power generation is about producing electricity with solar, wind and other sources of energy

Introduction of study activity is as follows. T is teacher, and S is student.

1) Activity 1: What's the situation?
T:I will share with you a story about Jaeho's family. Listen carefully and imagine what you'll do if you were Jaeho.
S:(Listens to the teacher.)
T:So what happened to Jaeho's family?

S:They couldn't use all the electronic appliances because of the high electric bill / They should save electricity / They have an issue with electricity , etc.
T:Let's check the situation Jaeho's family is in by using appliances together.
S:(Students use fan, lamp, radio and electric vehicles with limited electricity.)

2) Activity 2: Make smart grid
T:Let's think about the concept of smart grid and take advantage of it to solve problems.
S:(Students simulate simple smart grid with limited electricity consumption for fan, lamp and electric vehicle.)

3) Activity 3: Check together
T:Let's check if the simple smart grid works properly.
S:(Students confirm operation of the smart grid system.)

4 Result of Implementing Education Program and Analysis of Effectiveness

The questionnaire comprises 5 questions developed based on Likert Scale and 5 descriptive questions, and the result of implementing the education system is shown in (Table 1).

Table 1. Result of implementing the education

Subject	Frequency (Person)	Average	Standard Deviation
Level of difficulty of the class and learning materials	10	4.20	.75
Level of words and sentences used	10	4.20	.75
Interest in and satisfaction with learning materials and the class	10	4.70	.48
Preference to learning through the program in the future	10	4.60	.66
Degree of program's effect in connection with other school subjects	10	4.40	.66

According to the survey, the level of difficulty of the class and learning materials seemed appropriate as it showed 4.20 in average, so did the level of words and sentences used as it also showed 4.20 in average. The interest in and satisfaction with learning materials and the class showed 4.70 in average, meaning that respondents were quite interested in and satisfied with learning materials and the class. The preference to learning through the program in the future was 4.60, meaning that they highly prefer learning in STEAM program in the future and the degree of program's effect in connection with other school subjects was 4.40, meaning that the effect is high.

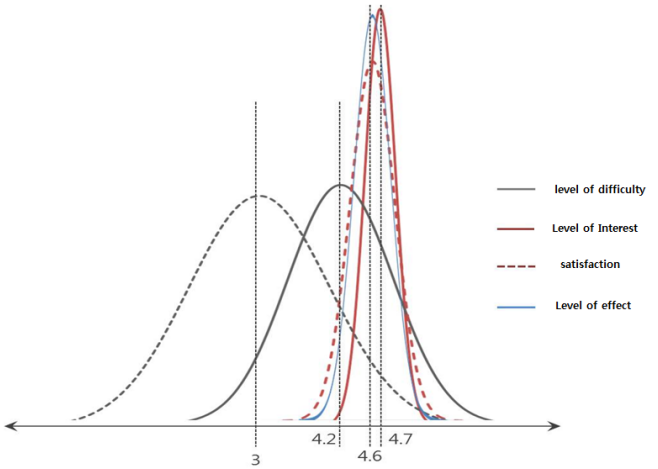


Fig. 1. Analysis of education level

Answers to descriptive questions were mostly positive about the program that implemented convergent education integrated with real life and new IT, and respondents were most satisfied with the 6th class, "Making Green Home", which integrated science, art, and Korean literature.

5 Conclusion and Extended Suggestions

STEAM education must have 5 domains converge on a single class in the course of solving problems instead of simply and mechanically putting them together. Also, learners concentrate in class more when it is about the issues related with real life and when the topic of study is exciting and interesting. In other words, STEAM education on IT is effective in fostering the convergent thinking and problem solving capabilities of students, and smart grid can enables effective STEAM education for it attracts students' interest and provides the problematic situation for students to develop the habit of convergent thinking.

Students can develop knowledge and view on IT while resolving problematic situations with an interest in STEAM education and smart grid through the STEAM education program on IT theme, particularly smart grid, suggested in this study. Furthermore, students, by themselves, can develop self-directed learning capabilities and problem solving capabilities by planning, carrying out, and evaluating the problem solving process.

This paper applied STEAM education program on IT theme, particularly smart grid, in a short period of time that it could only confirm the elements that exhibited changes in a short period of time which was not enough to conform problem solving capabilities or advanced thinking skills such as creativity. Such a limit can be

overcome by examining changes in students through follow-up studies after a certain period of time passes and extending and applying the curriculum according to the result of examination and also by developing many more STEAM education programs and learning materials on IT theme to establish STEAM education system in Korea.

Acknowledgments. This paper is extended from the SecTech/CA/CES-CUBE 2012, CCIS 339 'Development of Computer, Math, Art Convergence Education Lesson Plans Based on Smart Grid Technology (2012.09)'. The Corresponding author is Namje Park (namjepark@jejunu.ac.kr).

References

1. Park, N., Ko, Y.: Computer Education's Teaching-Learning Methods Using Educational Programming Language Based on STEAM Education. In: Park, J.J., Zomaya, A., Yeo, S.-S., Sahni, S. (eds.) NPC 2012. LNCS, vol. 7513, pp. 320–327. Springer, Heidelberg (2012)
2. Ko, Y., An, J., Park, N.: Development of Computer, Math, Art Convergence Education Lesson Plans Based on Smart Grid Technology. In: Kim, T.-H., Stoica, A., Fang, W.-C., Vasilakos, T., Villalba, J.G., Arnett, K.P., Khan, M.K., Kang, B.-H. (eds.) SecTech, CA, CES3 2012. CCIS, vol. 339, pp. 109–114. Springer, Heidelberg (2012)
3. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
4. Park, N.: Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
5. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
6. Park, N.: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing 8(4), 205–219 (2011)
7. Park, N., Kim, Y.: Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 193–202. Springer, Heidelberg (2010)
8. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 142–152. Springer, Heidelberg (2010)
9. Park, N.: Customized Healthcare Infrastructure Using Privacy Weight Level Based on Smart Device. In: Lee, G., Howard, D., Ślęzak, D. (eds.) ICHIT 2011. CCIS, vol. 206, pp. 467–474. Springer, Heidelberg (2011)
10. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
11. Park, N.: The Implementation of Open Embedded S/W Platform for Secure Mobile RFID Reader. The Journal of Korea Information and Communications Society 35(5), 785–793 (2010)

Security Enhanced Unlinkable Authentication Scheme with Anonymity for Global Mobility Networks

Youngseok Chung[1,2], Seokjin Choi[1], Youngsook Lee[3], and Dongho Won[2,*]

[1] Electronics and Telecommunications Research Institute, Korea
[2] Department of Computer Engineering, Sungkyunkwan University, Korea
[3] Department of Cyber Investigation Police, Howon University, Korea
{yschung11,choisj}@ensec.re.kr, ysooklee@howon.ac.kr,
dhwon@security.re.kr

Abstract. Recently, Chung, Lee, and Won [1] proposed an improved authentication scheme with anonymity which remedies security faults showed by Youn, Park, and Lim [2]. Their improved scheme guarantees anonymity, but does not provide unlinkability. In their scheme, it is possible for attackers to know particular sessions, that have already been occurred several times, are originated by one same user. In this paper, we propose an unlinkable authentication scheme with anonymity by modifying Chung et al.'s scheme. Our scheme provides not only anonymity and security as the previous scheme does, but also unlikability against malicious mobile users. Since proposed scheme still uses only low-cost functions, it is suitable for mobility networks.

Keywords: anonymity, linkability, authentication, mobility network.

1 Introduction

With the increase of mobile users who take roaming services, maintaining users' sensitive information like an identity safely is becoming more important. This is because no one wants the invasion of his/her privacy caused by disclosing the identity. Also, the authentication schemes suitable for battery-powered mobile devices in global mobility networks are becoming more essential in terms of providing efficiency.

Recently, many authentication schemes with anonymity have been proposed for mobility networks. In addition, several attacks against the proposed schemes and countermeasures have been also presented. Chung et al. categorized the previous schemes into two groups: schemes using high-cost functions such as asymmetric and symmetric cryptosystems and schemes using low-cost functions such as one-way hash functions and exclusive-OR operations. They focused on the latter.

Zu and Ma [3] proposed an authentication scheme with anonymity using high-cost functions, and some proofs of weaknesses and improvements have been followed [4-7]. On the other hand, Chang, Lee, and Chiu [8] firstly proposed an enhanced authentication by basing on low-cost functions. After that, Youn el al. presented that Chang et al.'s

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_80,

scheme not only fails to provide the anonymity against passive adversaries and malicious mobiles users but also security against known session key attacks and side channel attacks without any countermeasures. Chung et al. proposed an improved anonymous authentication scheme to remedy vulnerabilities presented by Youn et al.

In this paper, we propose a security enhanced unlinkable authentication scheme with anonymity for global mobility networks. Basically, our scheme has a structure similar that of Chung et al.'s and uses only one-way hash functions and exclusive-OR operations as the previous scheme dose. Therefore, our scheme is still suitable for mobility networks.

The remainder of this paper is organized as follows. In Section 2, we review Chung et al.'s scheme. The proposed scheme is presented in Section 3. In Section 4, we analyze the security of our scheme. Finally, a concluding remark is given in Section 5.

2 Review of Previous Works

In this section, we review Chung et al.'s scheme. Table 1. contains notations used in Chung et al.'s scheme. These terms are also applied in the same way in this paper.

Table 1. Notations

Notations	Descriptions
MN	A mobile user
HA	The home agent of a mobile user
FA	The foreign agent of a foreign network
x	A common secret key for all legitimate users
n_X	A nonce generated by entry X
ID_X	The identity of an entry X
PW_{MN}	A password of MN
x_{MN}	A secret key for MN generated by HA
r_{MN}	A random secret parameter for MN generated by HA
h(.)	A collision free one-way hash function
$\|$	A concatenation
$\oplus$	A XOR operation

There are three phases in Chung et al.'s scheme: registration, authentication, and session key establishment. MN, HA, and FA are involved in these phases. It is assumed that each FA and HA share a long-term common secret key K_{FH} established by using key agreement method, such as the Diffie-Hellman key agreement protocol.

2.1 Registration Phase

In the registration phase, a mobile user MN who wants to register his/her home agent HA submits ID_{MN} and the selected password PW_{MN}. Then, HA computes the virtual identity $VID = h(ID_{MN}||r_{MN})$ and $R = VID \oplus PW_{MN}$ after generating x_{MN} and r_{MN} for him/her only. HA stores $(ID_{MN}, r_{MN}, VID, h(x_{MN}))$ secretly and delivers a smart card containing $\{ID_{MN}, ID_{HA}, VID, R, h(x_{MN}), h(x), h(.)\}$ to MN through a secure channel.

2.2 Authentication and Session Key Establishment Phases

It is assumed that MN who roams into the foreign network visits FA and FA needs to authenticate MN through HA. MN, HA, and FA perform the following steps.

1. MN inserts his/her smart card into the device and enters PW^*_{MN}. $MN's$ smart card generates n_{MN} and calculates $C = (R \oplus PW^*_{MN}) \oplus n_{MN}$. MN sends a login message $m_1 = \{Login\ req., n_{MN}, ID_{HA}\}$ to FA for authentication.
2. Upon receiving m_1, FA records n_{MN}, generates n_{FA} and sends an authentication message $m_2 = \{Authenticaton\ req., n_{FA}, ID_{FA}\}$ to HA.
3. After receiving m_2, HA checks ID_{FA} to determine whether it is an ally. If the result is valid, HA generates n_{HA} and sends $m_3 = \{n_{HA}, ID_{HA}\}$ to FA.
4. After receiving m_3, FA sends $m_4 = \{n_{HA}, n_{FA}, ID_{FA}\}$ to MN.
5. Upon receiving m_4, MN generates $SID = VID \oplus h(h(x)||n_{HA})$, $V_1 = h(n_{HA}||C)$, $SK = h(h(x_{MN})||ID_{MN}||ID_{FA}||n_{MN}||n_{FA})$, $V_2 = h(SK||n_{HA})$, and $S_1 = h(n_{FA}||\ SID||V_1||V_2||n_{MN}$. And then MN sends $m_5 = \{SID, V_1, V_2, n_{MN}, S_1, ID_{HA}\}$ to FA.
6. After receiving m_5, FA uses n_{FA} with the received SID, V_1, V_2, and n_{MN} to compute $S^*_1 = h(n_{FA}||SID||V_1||V_2||n_{MN})$. If S^*_1 and S_1 are equivalent, FA computes $S_2 = h(K_{FH}||n_{HA}||SID||V_1||V_2||n_{MN})$ and sends $m_6 = \{SID, V_1, V_2, n_{MN}, S_2, ID_{FA}\}$ to HA to verify whether MN is legal.
7. After receiving m_6, HA checks ID_{FA} to determine whether it is an ally. Then, HA computes $S^*_2 = h(K_{FH}||n_{HA}||SID||V_1||V_2||n_{MN})$ to check whether $S^*_2 = S_2$. If the result is valid, the identity of FA is authenticated, and HA continues to check the validities of VID, PW^*_{MN}, and SK as follows:
   (a) HA computes $VID^* = SID \oplus h(h(x)||n_{HA})$ and retrieves a set of $MN's$ secret information $(ID_{MN}, r_{MN}, VID, h(x_{MN}))$ using VID^* as a search word to check whether $VID^* = h(ID_{MN}||r_{MN})$.
   (b) If the result is valid, HA computes $C^* = n_{MN} \oplus VID$ and $V^*_1 = h(n_{HA}||C^*)$ to check whether $V^*_1 = V_1$. Note that, the equivalence between V^*_1 and V_1 implies that PW^*_{MN} equals PW_{MN}.
   (c) If they are equal, HA computes $SK^* = h(h(x_{MN})||ID_{MN}||\ ID_{FA}||n_{MN}||n_{FA})$ and $V^*_2 = h(SK^*||n_{HA})$, and checks the validity of V^*_2.
8. Then, HA computes $K_1 = SK \oplus h(K_{FH}||n_{FA})$, $V_3 = h(ID_{FA}||h(x_{MN})||n_{MN})$, and $S_3 = h(K_{FH}||n_{FA}||K_1||V_3)$ and sends $m_7 = \{K_1, V_3, S_3\}$ to FA to inform that MN is a legal user.
9. With a message m_7, FA computes $S^*_3 = h(K_{FH}||n_{FA}||K_1||V_3)$ to check whether $S^*_3 = S_3$. Then FA computes $SK = K_1 \oplus h(K_{FH}||n_{FA})$, $K_2 = SK \oplus h(SK||n_{MN})$ and sends $m_8 = \{V_3, K_2\}$ to MN.
10. After receiving m_8, MN computes $V^*_3 = h(ID_{FA}||(x_{MN})||n_{MN})$ and checks where $V^*_3 = V_3$. If the result is valid, MN computes $SK^* = K_2 \oplus h(SK||n_{MN})$. If SK^* and SK are equal, MN is sure that FA also has an authenticated session key. Then, MN records the authenticated session key SK for future communications.

3 Proposed Scheme

In this section, we demonstrate an anonymous authentication scheme which provides unlinkability. It consists of four phases: registration, authentication, session key establishment, ID renewal phases. Authentication, session key establishment, and ID renewal phases are accomplished in one process.

3.1 Registration Phase

The registration phase is identical to that of Chung et al.'s scheme. Fig. 1. represents the registration process between MN and HA.

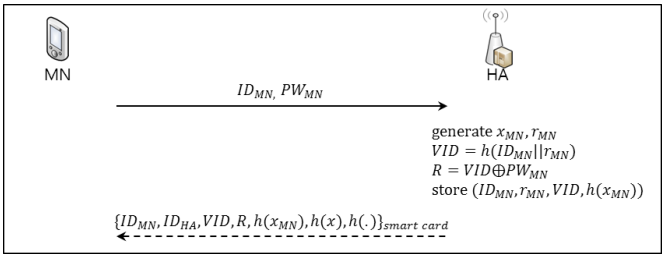


Fig. 1. Registration phase

3.2 Authentication, Session Key Establishment and ID Renewal Phases

Basically, the proposed scheme has similar assumptions and structures in the authentication and session key establishment phases of Chung et al's scheme. However, the ID renewal procedure is existed in the middle of the authentication and session key establishment phases in order to renew a user's ID. Fig 2. shows the modified phases of our scheme.

Step1 through Step 7 are the same as those of the previous scheme.

In step 8, after computing K_1 and V_3, HA generates a new random secret parameter $r_{MN}{}'$ and computes $VID' = h(ID_{MN}||r_{MN}{}')$. HA updates $(ID_{MN}, r_{MN}, VID, h(x_{MN}))$ to $(ID_{MN}, r_{MN}{}', VID', h(x_{MN}))$. New $r_{MN}{}'$ and VID' will be used to authenticate MN in the future. Then, HA computes the renewed ID $RID = VID' \oplus h(VID||h(x_{MN})||n_{MN})$, the hashing value $S_3 = h(K_{FH}||n_{FA}||K_1||V_3||RID)$ and sends a message $m_7 = \{K_1, V_3, S_3, RID\}$ to FA to inform that MN is a legal user.

After receiving m_7, FA computes ${S^*}_3 = h(K_{FH}||n_{FA}||K_1||V_3||RI$. If ${S^*}_3 = S_3$, FA computes the session key $SK = K_1 \oplus h(K_{FH}||n_{FA})$, $K_2 = SK \oplus h(SK||n_{MN})$ and sends a message $m_8 = \{V_3, K_2, RID\}$ to MN.

With a message m_8, MN computes ${V^*}_3 = h(ID_{FA}||(x_{MN})||n_{MN})$ and checks where ${V^*}_3 = V_3$. If they are equal, MN computes $SK^* = K_2 \oplus h(SK||n_{MN})$. If SK^* and SK are equal, MN is sure that FA also has an authenticated session key.

Finally, MN computes $VID' = RID \oplus h(VID||h(x_{MN})||n_{MN})$, $R' = VID' \oplus PW_{MN}$ and renews VID, R to VID', R'.

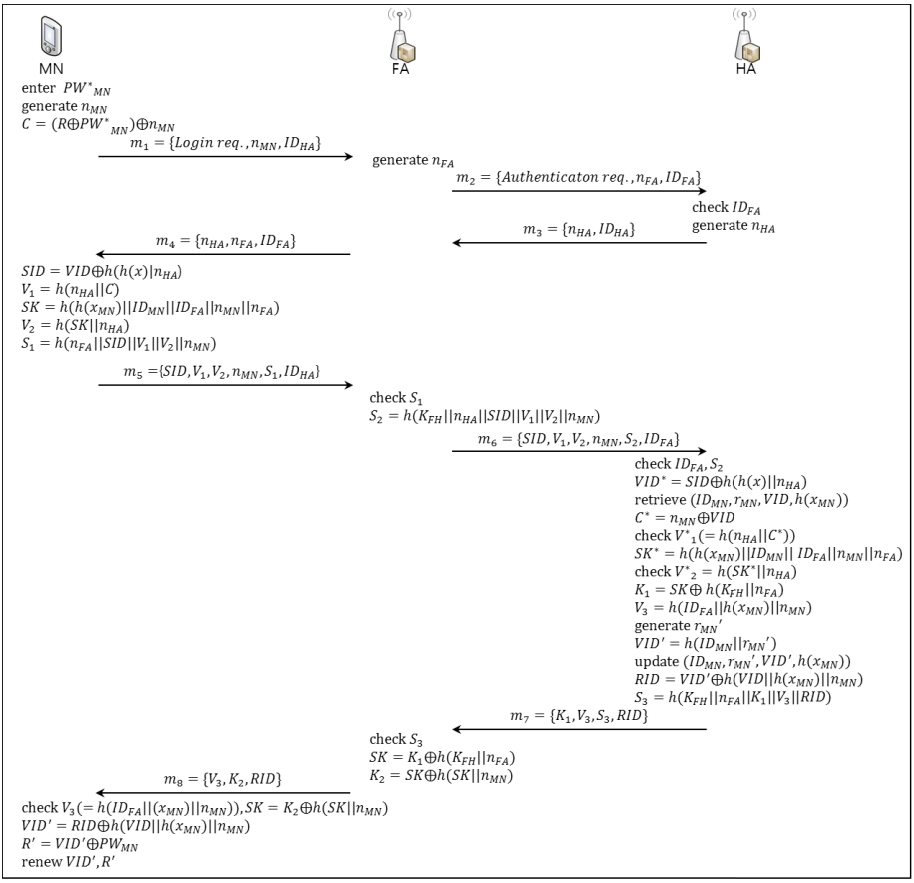


Fig. 2. Authentication, session key establishment and ID Renewal phase

4 Security Analysis

The proposed scheme still provides not only anonymity against passive adversaries and malicious mobile user but also security against known session key attacks and side channel attacks since it inherits the security attributes of Chung et al.'s scheme. Furthermore, it guarantees unlinkability that the previous scheme does not provide. It is assumed that there is a malicious user who possesses a valid smart card issued by HA. It is clear that he/she knows $h(x)$ which all legitimate users have in their smart cards. And he/she can easily eavesdrop n_{HA} and SID out of m_4 and m_5 respectively. Since $SI = VID \oplus h(h(x)||n_{HA})$ and VID always has a same value in Chung et al.'s scheme, a malicious user can trace the usage of VID of MN by capturing it

in every session. Namely, he/she can know the fact that the established sessions between MN and FA are belongs to one same user. On the other hand, $VID = h(ID_{MN}||r_{MN})$ is always renewed to $VID' = h(ID_{MN}||r_{MN}')$ in every session by ID renewal procedures in our scheme. Therefore, no one can know that to whom the particular session between FA and someone belongs.

5 Conclusion

In this paper, we presented the anonymous authentication scheme providing unlinkability. Unlinkability means no one can even know the fact that particular sessions are originated by one same user without knowing who is an actual user. By adding the ID renewal phase and keeping freshness of virtual identities, security enhancements are achieved in our scheme. In other words, users do not use the same virtual identity any more in every session. So it is impossible to recognize that whose session this is although attackers can keep watch specific user's session. Moreover, since the proposed scheme still uses only low-cost functions such as one-way hash functions and exclusive-OR operations, it is suitable to be applied to mobility networks.

References

1. Chung, Y., Lee, Y., Won, D.: Improved Authentication Scheme with Anonymity for Roaming Service in Global Mobility Networks. In: Park, J.J. (J.H.), Arabnia, H.R., Kim, C., Shi, W., Gil, J.-M. (eds.) GPC 2013. LNCS, vol. 7861, pp. 752–760. Springer, Heidelberg (2013)
2. Youn, T.Y., Park, Y.H., Lim, J.: Weaknesses in an anonymous authentication scheme for roaming service in global mobility networks. IEEE Communications Letters 13(7), 471–473 (2009)
3. Zhu, J., Ma, J.: A new authentication scheme with anonymity for wireless environments. IEEE Transactions on Consumer Electronics 50(1), 231–235 (2004)
4. Lee, C.C., Hwang, M.S., Lio, I.E.: Security enhancement on a new authentication scheme with anonymity for wireless environments. IEEE Transactions on Industrial Electronics 53(5), 1683-1687 (2006)
5. Wu, C.C., Lee, W.B., Tsaur, W.J.: A secure authentication scheme with anonymity for wireless communications. IEEE Communications Letter 12(10), 722–723 (2008)
6. Jeon, W., Kim, J., Nam, J., Lee, Y., Won, D.: An enhanced secure authentication scheme with anonymity for wireless environments. IEICE Transactions on Communications E95-B(7), 2505–2508 (2012)
7. Nam, J., Kim, S., Park, S., Won, D.: Security analysis of a nonce-based user authentication scheme using smart cards. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E90-A(1), 299–302 (2007)
8. Chang, C.C., Lee, C.Y., Chiu, Y.C.: Enhanced authentication scheme with anonymity for roaming service in global mobility networks. Computer Communications 32, 611–618 (2009)

A Feature-Based Small Target Detection System

Jong-Ho Kim[1], Young-Su Park[1], Sang-Ho Ahn[2], and Sang-Kyoon Kim[1,*]

[1] Department of Computer Engineering, Inje University, Gimhae, Gyeongsangnam-do 621-749, Republic of Korea
[2] Department of Electronic Engineering, Inje University, Gimhae, Gyeongsangnam-do 621-749, Republic of Korea
luckykjh@daum.net, pzworko@gmail.com, {elecash,skkim}@inje.ac.kr

Abstract. Existing small target detection systems generally use the difference image between a predicted background image and an original image. This method has two disadvantages. First, to predict the background image, the size of the structural element has to be carefully selected considering the size of small targets. Second, because of blurring, clutter such as clouds can occur around the edge of the background. To deal with these problems we propose a new feature-based detection system. The proposed method selects candidate pixels with Harris corner detector and then, again selects pixels that have a higher intensity than a threshold among the candidates. After labeling the selected candidates in order to obtain the number of pixels they have, the system decides which is a small target. In an experiment, our proposed method gave better results than the existing methods.

Keywords: Harris corner detector, New White Top-Hat, Labeling, Histogram.

1 Introduction

Concomitant with the development of scientific techniques, techniques and systems for the development of weapons have been making rapid progress. In particular, missiles and unmanned aerial vehicles (UAVs) that effectively strike targets over long distances have emerged as important threat elements, such that an effective counterstrategy is required.

Infrared Warning System (IRWS) and Infrared Searching and Tracking system (IRST) have been proposed as means of detecting small targets such as missiles early and to judge whether they are threats. In these systems, the difference image between a predicted background image and an original image is generally used to detect small targets from IR images that have a lot of clutter. The targets are assumed to occupy a couple of pixels.

Conventional methods utilize Max-Mean and Max-Median filters [1] and White Top-Hat(WTH) Transformation [2]. Max-Mean and Max-Median filters remove clutter and make a predicted background image from IR images while preserving the

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_81,

edges of clouds and structural backgrounds. After extracting a difference image using the background image, small targets are then detected from candidate targets with a threshold. However, the filters are not useful in cases where a target is on the edge of clouds or an input IR image is not clear. In addition, the performance of the filters is degraded when the shape of targets is irregular or a lot of clutter is distributed.

WTH transformation [2] also removes clutter and makes a predicted background image. This morphology-based method can detect targets in real-time because it is not a time consuming job. However, it also does not work well with a variety of pixel size targets. New White Top-Hat (NWTH) transformation [3], which uses several kinds of structural elements to get better background images and to cope well with the size of targets, has been proposed. However it is not flexible with regards to size because the available number of structural elements is limited. Moreover, one of its side effects is generation of clutter on the edges of clouds. Multi-structuring elements (multi-SEs) NWTH transformation [4] has been proposed to automatically determine the optimal size of structural elements. However, all these methods that use a difference image have difficulties dealing with the size of targets and clutter on the edges of clouds.

In this paper, we propose a new feature-based detection system to deal with the problems. The proposed method classifies pixels into two groups specifically, pixels for the background and pixels for the target using Harris corner detector. The detector extracts well corner features from images, so that it is able to find pixels corresponding to the boundaries of targets. More possible pixels are subsequently screened by a threshold of intensity, while almost all clutter is eliminated. Because the pixels can be a part of clouds or a target, a target is detected by labeling and finding the number of pixels of labeled areas.

The proposed method is not restricted by the size of the targets because it uses the finds pixels for edges of targets with Harris corner detector. In addition, it clearly divides images into areas for targets and backgrounds, which improves the detection performance by excluding clutter.

2 Small Target Detection System

2.1 The Main Structure of the System

Fig. 1 shows the main structure of the proposed small target detection system. First, it extracts corner features from input images using Harris corner detector. Next, it selects more plausible pixels that have high intensity and can be regarded as pixels for targets or backgrounds. The selected pixels then are labeled and areas that can be divided into targets and backgrounds generated. Finally, small targets are detected according to the number of pixels in the labeled areas with a size threshold.

Fig. 1. Main structure of the small target detection system

2.2 Feature Point Extraction

2.2.1 Harris Corner Detector

Harris corner detector is a popular interest point detector due to its strong invariance to rotation, scale, illumination variation, and image noise. Harris corner detector is based on the local auto-correlation function of a signal [5]; where the local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions. Our proposed method extracts corner features using the detector. The corner features consist of edges of targets and clouds, which are very useful in classifying images into interest areas such as targets and clouds. The detector algorithm is as follows:

Given a shift $(\Delta x, \Delta y)$ and a point (x, y), the auto-correlation function is defined as,

$$c(x,y) = \sum_{W} [I(x_i, y_i) - I(x_i + \Delta x, y_i + \Delta y)]^2 \tag{1}$$

Where $I(\cdot,\cdot)$ denotes the image function and (x_i, y_i) are the points in the window W (Gaussian1) centered on (x, y). The shifted image is approximated by a Taylor expansion truncated to the first order terms:

$$I(x_i + \Delta x, y_i + \Delta y) \approx I(x_i, y_i) + [I_x(x_i, y_i) I_y(x_i, y_i)] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{2}$$

where $I_x(\cdot,\cdot)$ and $I_y(\cdot,\cdot)$ denote the partial derivatives in x and y, respectively. Substituting approximation Eq. (2) into Eq. (1) yields

$$c(x,y) = \begin{bmatrix} \Delta x & \Delta y \end{bmatrix} \begin{bmatrix} \sum_{W} (I_x(x_i, y_i))^2 & \sum_{W} I_x(x_i, y_i) I_y(x_i, y_i) \\ \sum_{W} I_x(x_i, y_i) I_x(x_i, y_i) & \sum_{W} (I_x(x_i, y_i))^2 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{3}$$

The second matrix is $c(x, y)$ and degree of corner is detected. The discrimination of corner degree is determined by corner response function equations (4) [6]

$$R(x,y) = \det(C(x,y) - k[trace(C(x,y))]^2 \tag{4}$$

Threshold k is 0.04, corner response function $R(x, y)$ presents overall pixels. The tiny degree of corner response value is removed by threshold.

2.2.2 Selection of Pixels

In the pixels selection step, the pixels extracted by Harris corner detector are screened by an intensity threshold. Generally, a small target has a higher intensity than backgrounds and clutter because targets radiate a lot of heat. However, an IR image does not have a constant brightness value, which varies according to the kind of target, distance from the infrared ray sensor, and light scattering. Therefore, we screen

the pixels with adaptive thresholds. To automatically obtain the adaptive thresholds, we use the following method.

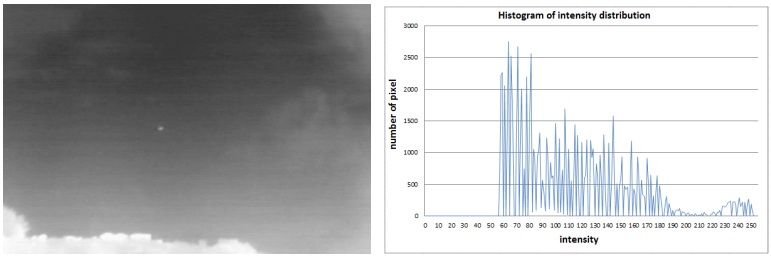
Fig. 2(b) shows the histogram of intensity distribution for the input image in Fig. 2(a). It shows that the intensity is not evenly distributed but rather biased to some values. The threshold is automatically obtained according to the distribution of intensity. Equation (5) is the method used to obtain the histogram of the intensity distribution [7]:

$$H(X_k) = n_k \tag{5}$$

Where X_k is the intensity of the k th intensity and n_k is the number of pixels that have the same value X_k . In this paper, the threshold of intensity T is determined by equation (6), which gives the median of the number of pixels in the histogram:

$$\begin{aligned} &IF\ \ N > (W \times H)/2\ \ Then \\ &T = k\ \ \ for \sum_{k=0}^{L-1} H(X_k) \end{aligned} \tag{6}$$

, where W is the width of the image and H is the height. Fig. 3 shows the threshold obtained. Fig. 4 shows that screening pixels and clutter are eliminated by the adaptive threshold. Fig. 4(a) shows the image that results after using Harris corner detector while Fig. 4(b) shows the image after clutter elimination.



(a) Input image (b) Histogram of intensity distribution

Fig. 2. Histogram of intensity distribution of an IR image

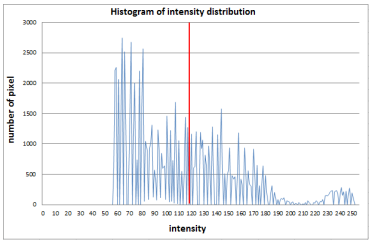
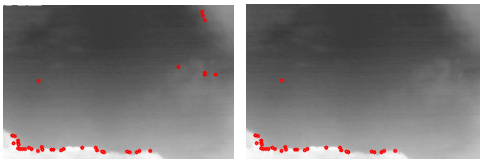


Fig. 3. An adaptive threshold

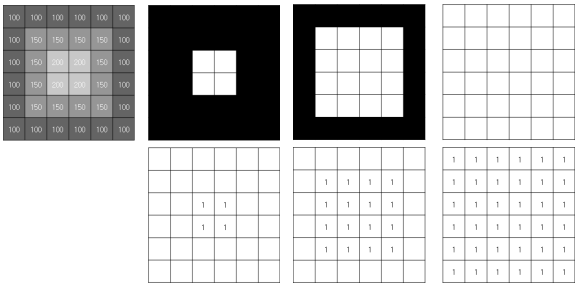


(a) (b)

Fig. 4. Screening pixels

2.3 Labeling and Making Areas

Screened pixels are labeled as a precursor to dividing them into two groups backgrounds and targets. Generally, binarization of an image is followed by labeling. However, original image information can be damaged when a fixed threshold is used in the binarization. For example, Fig. 5(a) shows how the results of binarization can differ according to the threshold used. In the labeling process, if a high threshold is used, targets or clouds can disappear.



(a) Gray Image (b) Threshold 190 (c) Threshold 140 (d) Threshold 90

Fig. 5. Labeling of binary images using thresholds

In this paper, the screened pixels are labeled using the pixel value of gray level without binarization to avoid information damage. Equation (7) shows how the neighbor in gray images is determined.

$$\begin{cases} Neighbor & : I(j)\times\omega_{Min} \le I(p_i) \le I(j)\times\omega_{Max} \\ Non-Neighbor: & otherwise \end{cases} \tag{7}$$

Where I is the intensity value, p_i is pixels, and i is the number of screened pixels. ω_{Min} and ω_{Max} are the minimum and maximum weights, respectively. In our experiments, we used 0.92 as the minimum weight and 1.3 as the maximum weight.

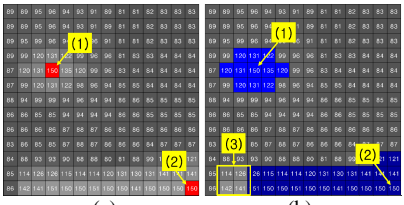
In traditional labeling methods [8], the overall pixels in an image are related. However, we perform the labeling operation only on the screened pixels. The actual size of a small target is very tiny because, for IR images, it is located a long distance

away from the camera. The size of a target is usually in ranges such as 3×3, 5×5, 7×7, 9×9, and 11×11. Thus, the search areas for the neighbor is limited to a size of 21×21. The strategy used to search for a neighbor is as follows:

• If a pixel among 4-connected neighbor pixels is not labeled, the pixel is labeled as a neighbor with the same label number.

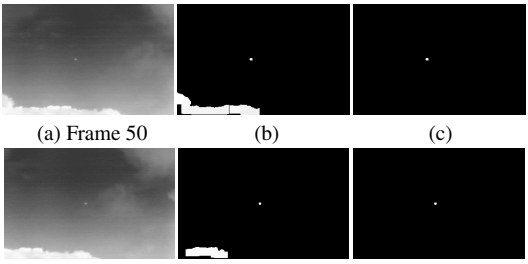
For instance, pixels (1) and (2), marked in red in the gray image of Fig. 6(a), are the screened pixels. Fig. 6(b) shows the result of our labeling. The pixels in area (3) are not included in area (2). Because of the limited search space, the larger neighbor does not need to be searched. Figs. 7(b) and 7(e) are the resultant images for the real images of Fig. 7(a) and 7(d), respectively.

In the making of areas, our labeling not only searches for only the screened pixels instead of all the pixels in the image but also limits the search areas that result from the tiny size of targets. Thus, our labeling improves traditional labeling methods and plays an important part in the performance of the entire detection system.



(a) (b)

Fig. 6. Search areas of the labeling: (a) screened pixels; (b) result of the labeling



(a) Frame 50 (b) (c)

(d) Frame 83 (e) (f)

Fig. 7. (a),(d) input images (b),(e) labeled image (c),(f) detected small targets from images (a), (d), respectively

2.4 Small Target Detection

The small targets are detected by the number of pixels in the labeled areas:

$$\begin{cases} Small\ Target : T_{Min} \le P(i) \le T_{Max} \\ Clutter \quad : \quad otherwise \end{cases} \tag{8}$$

In equation (8), $P(i)$ is the number of pixels in the i th labeled area. In this paper, threshold T_{Min} is 10 and T_{Max} is 150. Figs. 7(c) and 7(f) show the small targets detected from the real IR images of Figs. 7(a) and 7(d), respectively.

3 Experimental Results

Fig. 8 shows the simulation results for WTH transformation, NWTH transformation, multi-SEs NWTH transformation, and our proposed method, respectively. The experimental IR images had resolutions of 360×240 pixels. The IR images of the first and second row had a cloudy background and only one small target, respectively. The IR image of the last row had a cloudy background and only one small target in the cloud.

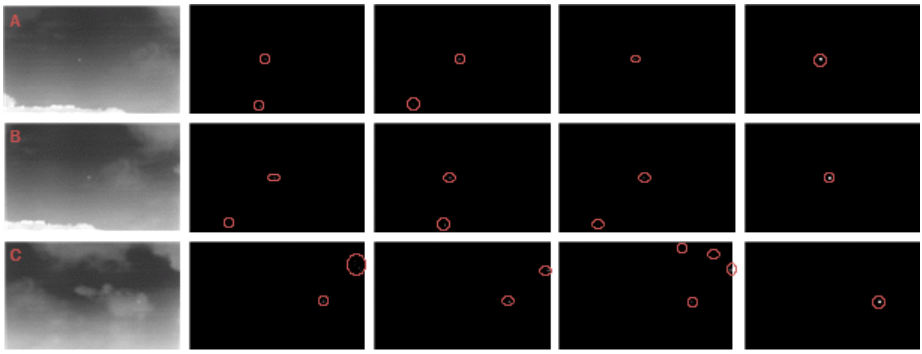


Fig. 8. (a) Input IR images; (b) WTH transformation; (c) NWTH transformation; (d) Multi-SEs NWTH transformation (e) Our proposed method

We analyzed each method in terms of processing time and correct detection rate. Table 1 shows processing times of each method for the real IR images in Fig. 8. Table 2 compares their results for correct detection rate.

The experimental results show that target detection ability of our proposed method is superior to that of WTH transformation, NWTH transformation, and multi-SEs NWTH transformation.

Table 1. Processing times

	WTH (s)	NWTH (s)	multi-SEs NWTH (s)	Proposed (s)
A	0.005	0.005	0.014	0.058
B	0.005	0.006	0.015	0.057
C	0.006	0.006	0.017	0.059

Table 2. Correct detection rate(%)

	Detection rate (%)			
	WTH	NWTH	multi-SEs NWTH	Proposed
Small target data	59.71	63.03	83.41	100.00

4 Conclusion

In this paper, we proposed a feature-based small target detection system from IR images. The system is flexible to the sizes of targets as it uses the corner features extracted with Harris corner detector. The clutter on the outside of targets and cloud regions are eliminated using a histogram of intensity distribution and adaptive thresholds. The clutter on the edges of clouds is also removed by our improved labeling technique. Experimental results show that our proposed method is more effective in detecting small targets than existing methods.

References

1. Deshpande, S.D., Er, M.H., Ronda, V., Chan, P.: Max-mean and max-median filters for detection of small-targets. In: Proc. SPIE, vol. 3809, pp. 74–83 (1999)
2. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, N.Y. (1982)
3. Bai, X., Zhou, F., Xie, Y.: New class of top-hat transformation to enhance infrared small targets. Journal of Electronic Imaging 17(3), 0305011–0305013 (2008)
4. Bae, T.W., Kim, B.I., Zhang, F., Kim, Y.C., Ahn, S.H., Sohng, K.I.: Recursive multi-SEs NWTH method for small target detection in infrared images. IEICE Electronics Express 8(19), 1576–1582 (2011)
5. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of Interest Point Detectors. International Journal of Computer Vision 37(2), 151–172 (2000)
6. Harris, C., Stephens, M.: A combined corner and edge detector. In: Fourth Alvey Vision Conference, pp. 146–151 (1988)
7. Park, G.H., Cho, H.H., Choi, M.R.: A Contrast Enhancement Method using Dynamic Range Separate Histogram Equalization. IEEE Trans. on Consumer Electronics 49(4), 1301–1309 (2003)
8. Angel, E.: Interactive Computer Graphics. Addsion Wesley Longman (1999)

A Small Target Detection System Based on Morphology and Modified Gaussian Distance Function

Jong-Ho Kim[1], Jun-Jae Park[1], Sang-Ho Ahn[2], and Sang-Kyoon Kim[1,*]

[1] Department of Computer Engineering, Inje University, Gimhae, Gyeongsangnam-do 621-749, Republic of Korea
[2] Department of Electronic Engineering, Inje University, Gimhae, Gyeongsangnam-do 621-749, Republic of Korea
luckykjh@daum.net, pjj0901@gmail.com, {elecash,skkim}@inje.ac.kr

Abstract. We propose a new small target detection system that detects small target candidates based on morphology operations and detects actual targets using a modified Gaussian distance function. To reduce clutter on the edges of clouds, a median filter is applied as preprocessing. Two kinds of images are calculated with closing and opening morphological operators, respectively. In the morphology operations, various sizes of structure elements are used to consider the sizes of targets and candidate targets are extracted from difference images between the two images in the closing and opening operations. With a modified Gaussian distance function, small targets are detected from the candidate targets. The proposed method is less sensitive to clutters than existing methods, and has a detection rate of 98%.

Keywords: IR Image, Small Target, Gaussian Distance Function, Top-Hat, NWTH.

1 Introduction

The small target detection method is morphology-based White Top-Hat(WTH) transformation[1]. WTH transformation operates subtraction between a source image and an image the performed opening operation. Because, the morphology-based method has simple calculation and is easy to realize, real-time processing is possible. However, it is vulnerable to images that have a lot of clutter because the resultant image is smooth. To overcome this disadvantage, New White Top-Hat(NWTH)[2],[3],[4] which improves structure elements to make them suitable shapes for targets and to reconstitute the method of operation, was proposed. NWTH transformation effectively removes much of the clutter distributed in a background; however, it cannot remove clutter that has a similar shape to a target among the background clutter because its structure elements are similar to the shape of the target. In addition, it cannot detect a target if the size of the structure elements is smaller than or the same with small target because it is sensitive to structure elements.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_82,

In this paper, we propose a robust small target detection method that overcomes the disadvantages of the existing methods. Our proposed method searches for small target candidates based on morphology operations, and detect the probability of a target by using a modified Gaussian distance coordinate function.

First, we remove clutter using a median filter for target detection. Next, closing and opening operations are performed by using various sized structure elements and target candidate pixels are obtained by means of a subtraction operation between the closing and opening operations. The target candidate pixels searched by the subtraction operation are clustered into candidate areas after labeling and the central location of the areas are calculated. Gaussian distance coordinate is modified in a source image by using the calculated centered location, and the area that is close to a small target is detected.

Because the target is removed by an opening operation in the resultant image of the subtraction operation and a target emerges clearly from the closing operation, an exact target is detected. In addition, because the resultant image is made after closing and opening operations of the image to be removed, the clutter emerging during the process of the subtraction operation is less than that of existing methods. Thus, it is more robust than the existing methods. Further, it reduces the false detection rate because it searches for candidate areas by probability using Gaussian distance coordinates.

2 Small Targets Detection System

2.1 System Structure

Fig. 1 shows the main structure of our small targets detection system.

First, we remove clutter included in the IR image using a median filter. Next, structure elements of various size are set according to the size of the target, and by means of a subtraction operation, resultant images are made using morphology based closing and opening operations. Candidate pixels are then selected from the resultant images of each subtraction operation. After which, the candidate pixels are labeled and clustered into candidate areas and centered coordinate candidates are detected. Finally, the small target is detected from the candidate areas using a proposed Gaussian distance function.

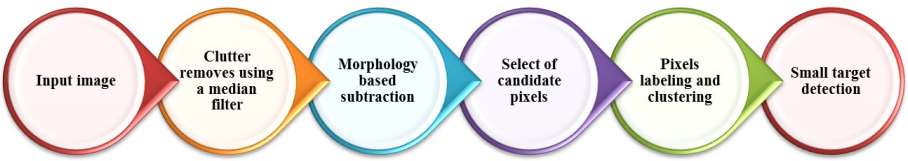


Fig. 1. Main structure of the proposed small targets detection system

2.2 Proposed Method

2.2.1 Clutter Removal using a Median Filter

The image from the infrared ray sensor includes a lot of clutter. Therefore, we remove clutter from the image using a median filter in a preprocessing phase.

2.2.2 Morphology Based Subtraction

After preprocessing using the median filter, we perform a morphology based subtraction operation. The morphology operators include dilatation $(f \oplus k)$, erosion $(f \otimes k)$, opening $(f \circ k)$, and closing $(f \bullet k)$. Eqs. (1)~(4) are basic morphology operation equations. The input image $f(x, y)$ consisting of gray-scale is presented by using structure elements $k(m, n)$.

$$f \oplus k = \max(f(x-m, y-n) + k(m, n)) \tag{1}$$

$$f \otimes k = \min(f(x+m, y+n) - k(m, n)) \tag{2}$$

$$f \bullet k = (f \oplus k) \otimes k \tag{3}$$

$$f \circ k = (f \otimes k) \oplus k \tag{4}$$

$$WTH = f(x, y) - [f \circ k](x, y) \tag{5}$$

A small target is brighter than the pixels in the surrounding regions, but it is small and has an ambiguous shape. Thus, when the opening operation is performed, the bright side of the image disappears, and when the closing operation is performed, the dark side of the image disappears. Our proposed morphology based subtraction operation performs the opening operation and closing operation not on the source image but on the resultant image from the median filter. The resultant image from the morphology based subtraction operation result image is obtained using Eq. (6) and is a marked subtraction operation from the next content.

$$S(x, y) = [f_{med} \bullet k](x, y) - [f_{med} \circ k](x, y) \tag{6}$$

$S(x, y)$ is the subtraction operation resultant image. $f_{med}(x, y)$ is a clutter image that is removed using the median filter. The sizes of small targets rang from 3×3 to 9×9. Thus, when structure elements are not appropriate for the size of a target, the misdetection rate increases. In other words, it is difficult to detect a target using structure elements of a fixed size. Therefore, in this pater, we set up structure elements with three size after considering the size of the small target. Each subtraction operation resultant image $S_{a \times a}$ $S_{b \times b}$ $S_{c \times c}$ is made about each of these structure elements.

2.2.3 Detection of Candidate Pixels

Because our proposed subtraction operation obtains the difference between pixels using the property of the closing operation and opening operation, the pixels of the small target are bright, while the pixels in its background are dark. However, it is not easy to discover the candidate small target pixels among the many pixels. In this paper, small target candidate pixels $\text{target}_{k\times k}$ are detected using Eq.(7) from the subtraction operation resultant image among the structure elements:

$$\text{target}_{k\times k}(x, y) \geq \alpha \times \max_{k\times k} \tag{7}$$

Where k is a, b, c (that is the sizes of the structure elements used in performing the subtraction operation): $\max_{k\times k}$ is the size of the brightest pixel value of an $S_{k\times k}$ resultant image; α is a parameter to extract the pixel that has the close brightness to $\max_{k\times k}$ and uses a value in the range $[0.1 \sim 1.0]$ in this paper. Pixels whose value is more than $\alpha \times \max_{k\times k}$ are candidate pixels. While those with value less than $\alpha \times \max_{k\times k}$ are considered background pixels. Because the more α the fewer the number of candidate pixels when discrimination a size of α, the exact shape of a small target doesn't realize. When α is too small, the brightest value realizes on the boundary line like a cloud and it occurs to miss a target.

2.2.4 Detecting Centered Coordinate of Candidate Region

The small target has a Gaussian distribution whose center is the brightest while the surrounding pixels are dark. To detect a target using the Gaussian distance function among the candidate pixels that have Gaussian distribution, preprocessing is needed.

Because we are detecting targets using candidate pixels, we need to know the location of candidate pixels of the source image. The centered coordinate about the location of the candidate pixels is obtained with the result of the subtraction operation. Checking the pixels all at once after clustering near the neighborhood pixels is faster than checking them one by one. Neighborhood pixels are thus clustered into a candidate area by using a labeling method.

The calculation of the centered coordinate of candidate targets precedes a binary according to candidate pixels and a background pixel that is determined from the subtraction operation resultant image. Candidate pixels are clustered into a candidate area using an algorithm based labeling method from a binary image. In the labeling process, the coordinates of the top, bottom, left, and right of each candidate region is obtained, and the centered coordinate is calculated using these four coordinates. Although the centered coordinate of the candidate areas is used in the source image, there is still a problem. Because the opening operation and the closing operation of the morphology operation operate from the left upper part of an image, the resultant image itself moves to the left upper part.

The resultant image from the morphology based subtraction operation is cut partially according to structure elements. Hence, because the centered coordinate is

calculated in a subtraction operation result image from the morphology operation, a process to shift the centered coordinate into a correct coordinate is needed. The method to calculate the correct coordinate is given in the form of Eq. (8):

$$\begin{aligned} x &= [(left + k - 2) + (right + k - 2)]/2 \\ y &= [(top + k - 2) + (bottom + k - 2)]/2 \end{aligned} \tag{8}$$

2.2.5 Probability Target Detection

Probability target detection uses a transformed Gaussian distance function and consider a target property that has a Gaussian shape. The existing Gaussian distance function is Eq. (9):

$$G_d(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[(x-m)^2 + (y-n)^2]\right) \tag{9}$$

When small target is detected using the existing Gaussian distance function, the importance of the pixel value of the small target becomes higher than the pixel value of its surroundings. Therefore, a modified Gaussian distance function such as that depicted in Eq. (10) is needed to decrease the importance of the pixel value of the target and to increase the value of the surrounding pixels.

This modified Gaussian function is defined by Eq. (10):

$$IvG_{k\times k}(x, y) = 1 - \lambda \cdot \exp\left(-\frac{1}{2\sigma^2}[|x-m| + |y-n|]\right) \tag{10}$$

Where α^2 is one and $m \times n$ is the centered coordinate of the structure elements. k represents the structure elements used to perform the subtraction operation. λ is a parameter representing distance weight and uses a value between [0.1] and [1]. The higher the value of λ is, the more importance is applied to the value of the surrounding pixels. If the importance of the center value of the Gaussian distance function proposed as Eq. (10) is zero, the meaning of the target central pixel disappears because the importance of the target central value disappears.

Therefore, $\frac{1}{2\pi\sigma^2}$ is replaced by λ to control the size of the central value of the Gaussian distance function, a method to adjust the importance of the distance using λ is selected, and standard normal distribution regarding σ^2 into one is used.

The method of detecting a small target using the proposed Gaussian distance function is illustrated Fig. 2. The location of the candidate regions in the source image is checked using the centered coordinate of the candidate area obtained in the labeling process. The candidate area location of the source image is then multiplied by $IvG_{k\times k}(x, y)$. At this time, the size of the structure elements used in the subtraction operation should be the same as that of k in $IvG_{k\times k}(x, y)$.

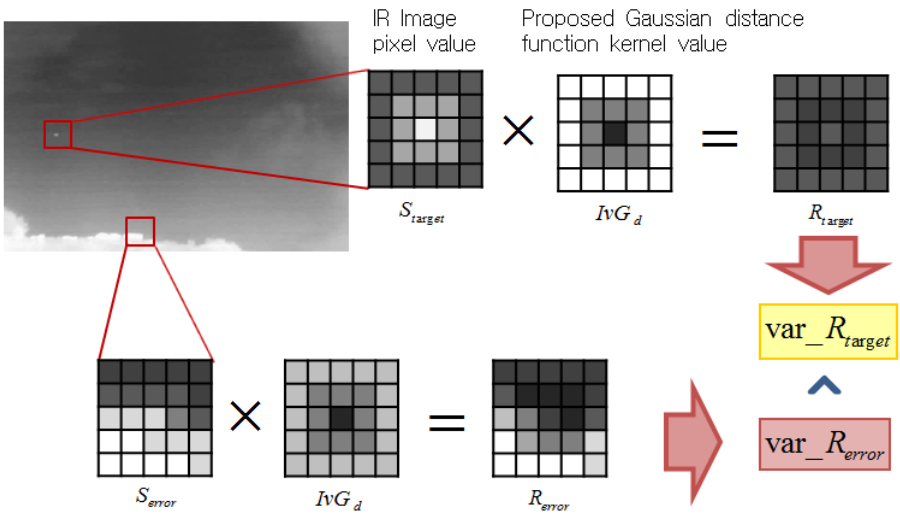


Fig. 2. Method of detecting a small target by using structure elements of the proposed Gaussian distance function

Structure elements R_{target} in Fig. 2 show that the importance of the small target pixel value is small and the importance of the background pixel value uses the brightness of the source image as is. Because the background is dimmer than the target pixel and the brightness value of the target is applied very little, the variance gap is small in cases where the candidate area is a target.

On the other hand, R_{error} is a cast that mistakes the boundary line for a candidate.

Because the importance of the central pixel brightness value of the structure elements decreases and the brightness of the source image is used as the pixels for the surrounding brightness, the variance gap is larger than R_{target}. In other words, the variance of R_{target} is less than that of R_{error}.

Therefore, the area that has the lowest variance value among the candidate areas according to the proposed Gaussian distance function becomes the target. Further, the structure elements that have the smallest R_{target} variance size among $a \times a$, $b \times b$, and $c \times c$ size structure elements that are used in the subtraction operation, are selected as the optimum size for the structure elements.

3 Experiment and Analysis of Results

The image data used in this experiment comprised 211 images of size 360×240 and single small targets in midair shot from the ground using a moving IR camera. In Fig. 3, the image in the first lines is the 37th among 211, the image of second lines is the 125th, and the image in the third lines is the 182nd. The 37th image among the 211 images has a bright cloud background and an obvious small target.

An image with the number 125 represents an image that has a highly complex background due to clouds. Such an image represents a situation in which a small target is hidden hides behind clouds.

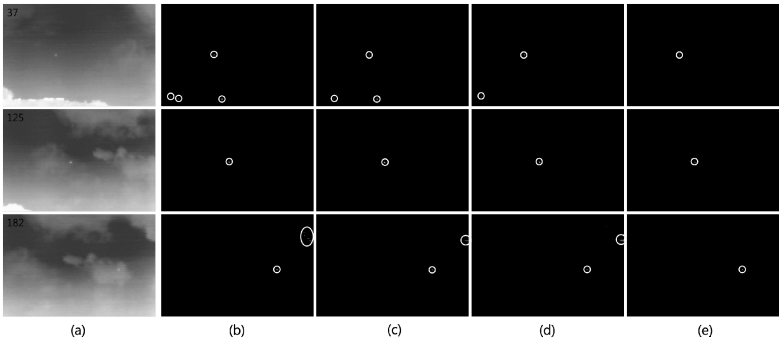


Fig. 3. (a) Input image, (b) WTH operation result image, (c) NWTH operation result image, (d) Multilayered structured NWTH operation result image, (e) Proposed algorithm result image

Figs. 3(b)~(c) are the result of using 5×5 structure elements. Fig. 3(d) results from multi-structure elements with initial condition a=5, b=9, c=13. The proposed algorithm Fig. 3(e) used a=5, b=7, c=9 structure elements and λ=0.7 in the Gaussian distance function. Further, Figs. 3(b)~(e) are the a result of using $0.7\times\max_{k\times k}$ to select candidate pixels. Figs. 3(b)~(d), 37 and 182 resultant images, have a mistake in that not only small targets but also boundary lines are detected as a target.

However, in Fig. 3(e) it can be seen that a small target that is accurate in probability among the small target area and the cloud boundary area is detected using the proposed modified Gaussian distance function.

Table 1 compares the resulting processing time and detection rate of the existing methods to those of our proposed method. Our proposed method has a longer processing time than all the existing methods; however, it has a higher detection rate than all of them.

Table 1. Processing times and detection rates

	WTH	NWTH	Multi-SEs NWTH[5]	Proposed
Processing time (clock/sec)	0.005	0.005	0.014	0.058
Detection rate(%)	61.16	63.50	83.41	98.57

The detection time is the measured time taken to detect one image. The detection of small targets accurately among the 211 images without mistakes is the detection rate.

4 Conclusion

In this paper, we proposed a detection method that detects small target candidates based on morphology operation and detects actual targets using a modified Gaussian distance function

Because the morphology operation should be performed on an image from which clutter has already been removed by a median filter, structure elements that are less sensitive than existing morphology operations are needed. Further, because morphology operations are performed using structure elements, structure elements that are suitable for a target are needed. The proposed method can detect targets more accurately because it takes into consideration the shape of a target using the Gaussian distance function. However, a disadvantage of our proposed method is the fact that its processing time is longer than that of existing methods due to the large amount of calculation it dose.

To solve these problems, a method that reduces the scanning area after predicting the target location is being investigated.

References

1. Gonzalez, C., Woods, R.: Digital Image Processing, 3rd edn. Prentice-Hall, Upper Saddle River (2008)
2. Bai, X., Zhou, F., Xie, Y.: New class of top-hat transformation to enhance infrared small targets. Journal of Electronic Imaging 17(3), 0305011–0305013 (2008)
3. Bai, X., Zhou, F., Xie, Y.: Analysis of new top-hat transformation and the application for infrared dim small target detection. Pattern Recognition 43, 2145–2156 (2010)
4. Bai, X., Zhou, F., Xie, Y.: Infrared small target enhancement and detection based on modified top-hat transformations. Computers and Electrical Engineering 17(3), 0305011–0305013 (2010)
5. Bae, T.W., Kim, B.I., Zhang, F., Kim, Y.C., Ahn, S.H., Sohng, K.I.: Recursive multi-SEs NWTH method for small target detection in infrared images. IEICE Electronics Express 8(19), 1576–1582 (2011)

Using Hardware Acceleration to Improve the Security of Wi-Fi Client Devices

Jed Kao-Tung Chang and Chen Liu

Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY
{jchang,cliu}@clarkson.edu

Abstract. As mobile devices prevail, communications security has become a critical and popular topic in recent years. For example, when a mobile device accesses Wi-Fi, the data communicated with the Wi-Fi access point may be encrypted to provide extra security. However, on a mobile device with limited energy budget, data encryption/ decryption often imposes high pressure on the battery life. In this paper, we review how a recently introduced Extremely Heterogeneous Architecture (EHA) can potentially be used to improve performance and energy efficiency of data encryption/decryption on mobile devices.

Keywords: Wi-Fi, Extremely Heterogeneous Architecture (EHA), cryptography, hardware acceleration, performance analysis, hotspot function.

1 Introduction

In recent years, mobile devices (Android, iPhone, and Windows Phone) have prevailed and are taking over desktop computers as the primary personal computing devices. Under this background, communication security has become a critical and popular topic. For example, when a mobile device accesses Wi-Fi, the data communicated with the Wi-Fi access point may be encrypted to provide extra security. However, on a mobile device with limited energy budget, data encryption/decryption often imposes high pressure on the battery life.

Under this constraint, users can choose to sacrifice the battery life to get extra security, or they can choose to sacrifice their data security in return of longer battery life. Fortunately, there is a recently developed hardware architecture that may be used to address this problem. The Extremely Heterogeneous Architecture (EHA), recently developed by Tang et al. [1-9], proposes to design specialized hardware accelerators to improve the performance as well as energy efficiency of the target workloads. This architecture has been used to accelerate many different workloads, such as XML parsing, software defined radio, garbage collection operations [10-15].

In this paper, we introduce this EHA architecture to the communication application, and we review how this new architecture can potentially be used to improve the performance and energy efficiency of the Wi-Fi encryption and decryption workloads.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_83,

2 EHA Architecture

In this section, we review the basic design concepts behind the EHA architecture. For more details about this architecture, please refer to the work done by Tang et al. [1-9].

As indicated in [20], the EHA consists of general-purpose cores and specialized cores, including:

- A heavy-weight high-performance core, which runs at high frequency and executes up to eight instructions each cycle;
- Multiple homogeneous light-weight low-performance cores, which run at low frequency and execute two instructions per cycle;
- Multiple hardware accelerators implemented using Application Specific Integrated Circuits (ASICs);
- Multiple reconfigurable accelerators implemented using Field Programmable Gate Arrays (FPGAs).
- Each of these cores will host a specific service.

In this design, the high-performance core acts as the "control core", duty of which includes scheduling, load balancing, housekeeping, as well as some basic computation tasks. The low-performance cores are the "execution cores", which get the tasks from the "control core" and execute them as scheduled. This type of "control core" can be used in mobile devices which apply Game theoretical and optimal approach [16]. Specialized hardware accelerators are dedicated for application acceleration and those targeted applications are normally frequently invoked or computationally intensive. To improve the performance of a particular workload, the EHA uses ASIC and FPGA to accelerate the execution of these programs. For applications that have great parallelism, the EHA uses the Graphics Processing Units (GPUs) for acceleration, the processing engines (PEs) of which can work concurrently to explore the potential parallelism. The EHA integrates all these heterogeneous cores into the same chip and they can communicate with each other via the high-speed on-chip interconnect network. As manufacturing technology progresses, a single chip can integrate more cores and more cache capacity. EHA can thus scale up well because the number of both generic and specific cores can be adjusted correspondingly.

3 Application of EHA on Encryption/Decryption Operations

In the past 30 years, more and more information is stored in digitized format due to the rapid development of computer and communication technology. However, without good protection scheme, data could be hacked and cracked by malicious intruders. A good security mechanism will be able to keep the secrecy and integrity of the information. Henceforth, data encryption/decryption has become an essential component for modern information exchange. For example, Wi-Fi access points usually use WPA2 [22] to perform security support.

However, these cryptography algorithms are often computation intensive, as many arithmetic and logic operations are executed in the encryption/ decryption process to

make data not easily cracked. Huge amount of data also are transmitted between the CPU and memory. Using a traditional general-purpose processor imposes very high computation and energy overheads, which would not be efficient for this scenario. Thus, the performance enhancement for the security component is crucial for mobile devices.

To address this issue, we can use the EHA architecture to expedite the cryptographic operations through hardware acceleration. Compared with many previous works [23-24] which focus on the algorithm design or the parallelization technique, one way to accelerate these encryption/decryption algorithms is to identify certain "hotspot functions" in the cryptographic algorithm, which consume a substantial amount of execution time of the specific algorithm. By using specialized hardware to accelerate these operations, we can greatly reduce the time and energy required to complete these expensive operations.

To this end, we [18] have collected nine benchmarks used widely in modern security systems: AES, 3DES, RC5, MD5, IDEA, SHA1, Blowfish, ECC and RSA. Rijndael's algorithm has been selected as AES due to its good balance in terms of speed, security, and flexibility. 3DES applied the DES three times to each data block and is widely adopted in banking information system and electronic payment industry. IDEA is used by PGP (Pretty Good Privacy) v2.0 to transmit message bodies. Blowfish provides a good encryption rate and its memory footprint is also very small.

ECC and RSA are public key algorithms. ECC is based on the algebraic structure of elliptic curves over finite fields. It is now popular due to the fact that it offers the same security level as offered by other contemporary algorithms at a shorter key length. RSA is suitable for encryption and digital signature and used in E-Commerce protocols. Although the execution of ECC and RSA are time-consuming, in these asymmetric algorithms each data can be encrypted or decrypted independently and the operations on these data can be performed in parallel. RC5, MD5, and SHA1 are all hash algorithms, which are used to verify the integrity of data blocks. RC5 is notable for its simplicity. What's more, the length of its secret key, word size, and number of rounds of computation can be varied. It is used in devices with restricted memory size such as smart cards. MD5 is widely used to assure if the transmitted file has arrived intact and to store passwords. SHA1 is often used in firewall, VPN, and IP security.

The approach of us towards hardware acceleration consists of two steps. The first step is to identify the hotspot functions. There are three aspects of the hotspot function we identified as worth considering.

Firstly, we chose a hotspot function with high execution rate. The execution rate is the percentage of the execution time belonging to hotspot function over the entire execution for each benchmark. We employed performance analyzer Vtune [25] and showed that seven out of nine benchmarks had hotspot functions with high execution rate. These are the software performance bottleneck and if they are implemented into hardware accelerator, the overall performance will be enhanced dramatically.

Secondly, the hardware cost to implement hotspot function is equally important. If the execution rate of a hotspot function is too low, implementing this function in hardware would make little sense since it has little impact on overall benchmark performance. In some cryptographic algorithms, the hotspot function is the entire

main process of the total algorithm, such as the encryption/decryption part. The hardware cost would be too high and much die area would be consumed for such cases. In this situation, we examined at a finer granularity by stepping into the hotspot functions to see if there is a "hot-line" or "hot-block" (basic block or several lines of code) whose size is small but still consumes a significant amount of time.

Finally, the hotspot points (function/block/code lines) should be concentrated into one or up to a small number of places in the benchmark. If several hotspot points are distributed all over the benchmark, even they have high execution rate, these hotspot points are needed to be implemented as many separate hardware accelerators. Issues such as data dependency will make the design complex, and a lot of interfaces between the host (control core) and accelerators (execution cores) will take more power and hardware cost. Thus a hotspot function with high execution rate, low hardware cost, and high concentration will be suitable for hardware acceleration. We found AES has three hotspot functions and other six benchmarks has one hotspot function or hot-block, all of which are of small size and consume a lot of execution time. These hotspot points qualified for the above three aspects of hardware implementation.

Next, we implemented the identified hotspot function/hot-block of the selected benchmarks into accelerators on the platform of Xilinx Virtex-5 FPGA board running at 100MHz. The hardware implementations achieved 30-100 folds performance improvement compared to the pure software implementation of the hotspot functions. Also they achieved the speedups of 2.9 for RSA, 4.5 for AES Encryption, and 5.9 for AES Decryption in terms of overall execution time of the cryptography algorithms. We measured the power and energy consumption and found the accelerator consumes much less energy than the general purpose processor. In some cases the difference can be up to two to three orders of magnitude. We also measured the hardware cost in terms of the number of hardware slices, flip-flops, and look-up tables. The overhead incurred by the accelerator is quite minimal. Compared to a very simple in-order MIPS processor design [21], the accelerator took only 1/100th to 1/1000th hardware cost of the MIPS processor.

To implement this acceleration technique into EHA architecture, we could integrate these hardware accelerators as ASIC/FPGA accelerators proposed in the EHA and offload all the Wi-Fi encryption/decryption workloads to the hardware accelerators to achieve performance and energy efficiency. The hardware cost and energy consumption investigation showed that the accelerators as "execution core" did not take a lot of energy consumption and die area of the EHA architecture.

4 Conclusions

In this paper, to address the challenging issue of performance and energy overheads of data encryption/decryption occurred for mobile devices when communicating with the Wi-Fi access point, we introduce a recently developed hardware computer architecture, the Extremely Hardware Architecture, to the communication application. We dive into the details of the EHA framework and discuss how this architecture can be used to reduce the performance overhead and improve the energy efficiency of encryption/decryption workloads.

In the near future, the EHA architecture can be extended to accelerate more networking workloads as well.

Acknowledgements. This work is partly supported by the National Science Foundation under Grant No. ECCS-1301953. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Tang, J., Liu, S., Gu, Z., Liu, C., Gaudiot, J.-L.: Acceleration of XML Parsing Through Prefetching. IEEE Transactions on Computers (in press)
2. Tang, J., Thanarungroj, P., Liu, C., Liu, S., Gu, Z., Gaudiot, J.-L.: Pinned OS/Services: A Case Study of XML Parsing on Intel SCC. Journal of Computer Science and Technology (in press)
3. Tang, J., Liu, S., Gu, Z., Li, X.-F., Gaudiot, J.-L.: Achieving middleware execution efficiency: Hardware-assisted Garbage Collection Operations. Journal of Supercomputing 59(3), 1101–1119 (2012)
4. Tang, J., Liu, S., Gu, Z., Liu, C., Gaudiot, J.-L.: Prefetching in Mobile Embedded System Can be Energy Efficient. IEEE Computer Architecture Letters 10(1), 8–11 (2011)
5. Tang, J., Liu, S., Gu, Z., Li, X.-F., Gaudiot, J.-L.: Hardware-Assisted Middleware: Acceleration of Garbage Collection Operations. In: Proceedings of the 21st IEEE International Conference on Application-Specific Systems Architectures and Processors (ASAP 2010), Rennes, France, pp. 281–284 (2010)
6. Tang, J., Liu, S., Gu, Z., Liu, C., Gaudiot, J.-L.: Memory-Side Acceleration for XML Parsing. In: Altman, E., Shi, W. (eds.) NPC 2011. LNCS, vol. 6985, pp. 277–292. Springer, Heidelberg (2011)
7. Liu, S., Tang, J., Wang, L., Li, X.-F., Gaudiot, J.-L.: Packer: Parallel Garbage Collection Based on Virtual Spaces. IEEE Transactions on Computers (October 2011), doi:10.1109/TC.2011.193
8. Liu, S., Tang, J., Deng, C., Li, X.-F., Gaudiot, J.-L.: RHE: A JVM Courseware. IEEE Transactions on Education 54(1), 141–148 (2011)
9. Tang, J.: Research on Performance Acceleration and Energy Efficiency for EHA Multicore Processor. Ph.D. Thesis, Beijing Institute of Technology, China (2012)
10. Liu, S., Gaudiot, J.-L.: Potential Impact of Value Prediction on Communication in Many-Core Architectures. IEEE Transactions on Computers (June 2009)
11. Liu, S., Wang, L., Li, X.-F., Gaudiot, J.-L.: Space-and-Time Efficient Parallel Garbage Collector for Data-Intensive Applications. International Journal of Parallel Programming (October 2010)
12. Liu, S., Eisenbeis, C., Gaudiot, J.-L.: Value Prediction and Speculative Execution on GPU. International Journal of Parallel Programming (December 2010)
13. Liu, S., Pittman, R.N., Forin, A., Gaudiot, J.-L.: Minimizing the Run-time Partial Reconfiguration Overheads in Reconfigurable Systems. Journal of Supercomputing (in press)
14. Liu, S., Pittman, R.N., Forin, A., Gaudiot, J.-L.: Achieving Energy Efficiency through Run-Time Partial Reconfiguration on Reconfigurable Systems. ACM Transactions on Embedded Computing Systems (in press)

15. Liu, S., Ro, W.W., Liu, C., Salas, A.C., Cérin, C., Gaudiot, J.-L.: EHA: The Extremely Heterogeneous Architecture. In: Proceedings of the 2012 International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN 2012), San Marcos, Texas, December 13-15 (2012)
16. Shi, Z., Beard, C., Mitchell, K.: Tunable Traffic Control for Multihop CSMA Networks. In: IEEE Military Communications Conference, MILCOM 2008, pp. 1–7 (November 2008)
17. Chaparro-Baquero, G., Zhou, Q., Liu, C., Tang, J., Liu, S.: Power-Efficient Schemes Via Workload Characterization on the Intel's Single-chip Cloud Computer. In: Proceedings of the Eighth Workshop on High-Performance, Power-Aware Computing (HPPAC 2012), in conjunction with IPDPS 2012, Shanghai, China (May 12, 2012)
18. Chang, J.K.-T., Liu, C., Liu, S., Gaudiot, J.-L.: Workload Characterization of Cryptography Algorithms for Hardware Acceleration. In: Proceedings of the 2nd ACM/SPEC International Conference on Performance Engineering (ICPE 2011), Karlsruhe, Germany, March 14-16 (2011)
19. Liu, C., Duarte, R., Granados, O., Tang, J., Liu, S., Andrian, J.: Critical Path Based Hardware Acceleration for Cryptosystems. International Journal of Advancements in Computing Technology (IJACT) 4(1), 438–452 (2012)
20. Duarte, R., Granados, O., Liu, C., Andrian, J.: Case Study on a Software Communications Architecture Component for Hardware Acceleration. In: Proceedings of the 4th International Conference on Ubi-media Computing (U-Media 2011), Sao Paulo, Brazil, July 3-4 (2011)
21. The eMips project, `http://research.microsoft.com/en-us/projects/emips/default.aspx`
22. Lashkari, A.H., Danesh, M.M.S., Samadi, B.: A survey on wireless security protocols (WEP, WPA and WPA2/802.11 i). In: 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2009. IEEE (2009)
23. Bielecki, W., Burak, D.: Parallelization Method of Encryption Algorithms. In: Advances in Information Processing and Protection, pp. 191–204 (2008)
24. Beletskyy, V., Burak, D.: Parallelization of the IDEA algorithm. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3036, pp. 635–638. Springer, Heidelberg (2004)
25. Intel Vtune, `http://software.intel.com/en-us/intel-vtune/`

An Anonymous Communication Scheme with Non-reputation for Vehicular Ad Hoc Networks

Ching-Hung Yeh[1], Meng-Yen Hsieh[2], and Kuan-Ching Li[2]

[1] Department of Computer Science and Information,
Engineering Far East University Tainan, Taiwan
chyeh@cc.feu.edu.tw
[2] Department of Computer Science and Information,
Engineering Providence University Taichung, Taiwan
{mengyen,kuancli}@pu.edu.tw

Abstract. Vehicular Ad hoc Network (VANET) is a kind of open wireless communication network and uses 802.11p protocol to interconnect vehicles and provides numerous services. Although it brings many of convenient applications but unlike traditionally wired networks are protected by several defenses such as firewalls and gateways, it may face a variety of security challenges such as security and privacy. The users need anonymous mechanisms to enable unlink ability but impartial third party requests non-reputation for accidents or certain events. In brief, the privacy preserving and non-reputation are contradictory. To overcome these flaws, an anonymous communication scheme with non-reputation for vehicular ad hoc networks is proposed. Our proposed scheme not only accomplishes anonymously communication between vehicle-to-vehicle and vehicle-to-roadside infrastructure for protecting privacy, but also achieves non-reputation function for identifying vehicle.

Keywords: vehicular ad hoc network, security, privacy, non-reputation.

1 Introduction

Along with the improvement of wireless communication technologies, Vehicular Ad-hoc Networks (VANETs) [1] which are one of special form of Mobile Ad-hoc Networks (MANETs) [2], has been used to improve the driving experience [3]. Vehicles with onboard units (OBUs) support interactions among nearby vehicles and roadside units (RSUs) such that provide a better way to collect dynamic traffic information with low cost and high accuracy than man-made.

VANET as the internet, a variety of security challenges always exists. A malicious attacker can eavesdrop and records messages transmitted by targeted vehicle. Then, the attacker can track the location of the vehicle and infer the private information of its driver and passengers from recorded messages. It is clear that privacy and non-reputation play a critical role in the communications of the VANET and need to be enhanced to thwart such scenarios. Significant progresses [4][5] have been made in secure network models and threats in VANETs. Some researches [6][7] described

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_84,

secure key, group key and certificate framework. But some of existing schemes [8][9] only focused on authentication with privacy preservation without non-reputation mechanism. Their solutions are relied on a huge number of anonymous keys which will be a problem when the revoked anonymous keys are large.

Concerns both privacy preservation and non-reputation are a prerequisite for deployment of VANET, we proposed an anonymous communication scheme with non-reputation for vehicular ad hoc networks. In our scheme, vehicles generates private/public key pairs by itself and a Regional Authority (RA) serves as a trusted third party in which generate vehicular pseudonym and certificates of vehicular pseudonym.

The remainder of this paper is organized as follows. A briefly related works are discussed in Section 2, and the proposed scheme is introduced in Section 3. The security analysis and evaluation results are explained in Section 4, and finally, the conclusion and the future works are drawn in Section 5.

2 Related Work

Raya et al. [10] proposed using public key infrastructure (PKI) for VANET security which investigated the privacy issue by a pseudonym-based approach with PKI and certificate authority (CA). In their scheme, each OBU stores a set of anonymous public/private keys to sign traffic messages and changes private keys periodically for avoiding being traced. Although the scheme achieves anonymous message authentication and conditional privacy preservation, but accompanies some disadvantages. First, each OBU needs to keep a large number of anonymous pair-wise keys. Second, a high search time is inevitable on looking up the long revocation list for any dispute certificate. Third, OBUs must suffer the overhead of anonymous keys being revoked and updated the certificate revocation list. Moreover, such PKI security frameworks need extremely cost for public key certificate and storage overhead to manage the Certificate Revocation Lists (CRLs).

GSIS [11] used group signatures for OBUs and identity-based signatures for RSUs to resolve the flaw of a large number of public key certificates. A message can be verified sending from a legitimate sender by its signature which signs by group signature scheme. However, although it has the advantage of shorter revocation list and easily updated, but each vehicle has to spend more time on message verification, in which is linear proportion to the number of revoked vehicles in the revocation list. This solution is not feasible for time-aware messages.

C.T. Li et al. [12] proposed a secure and efficient communication scheme, which utilizes non-interactive ID-based public-key cryptography, blind signature, and one-way hash chain. But, it does not take the non-repudiation into account. These security solutions using identity-based cryptography in VANETs strongly depend on the infrastructure to perform private/public key pair assignment and pseudonym generation. The key generation infrastructure can abuse its access ability, which implies the key escrow problem. Briefly, the non-repudiation, private preserving does not be strongly guaranteed.

3 Anonymous Communication Scheme with Non-reputation

Intuitively, there is a Region Head-end (RH) located at the top layer, which is composed of a RA, application servers and necessary equipment. We assume that the RH is secure, which be managed and protected by legitimate organization so that network entities communicate with RIH is safe. For eased presentation, the notations throughout this paper for our proposed scheme are listed in Table 1.

Table 1. Notations and Descriptions

Notations	Descriptions
RA	A regional authority
G	A cyclic additive group
G_T	A multiplicative group
q	The order of groups G and G_T
P	The generator of the cyclic additive group G
$\hat{e}$	A bilinear map $G * G \rightarrow G_T$
PK_i	The public key of entity i
$PK_{i;j}$	The j-th public key of entity i
Sk_i	The private key of entity i
$SK_{i;j}$	The j-th private key of entity i
$H_1()$	A MapToPoint hash function, an arbitrary string to G, $H_1 : \{0,1\}^* \rightarrow G$
$H_2()$	A MapToPoint hash function that maps some n to G_T, $H_2 : \{0,1\}^n \rightarrow G_T$
h()	A strong one-way hash function
RID_i	The real identity of entity i
PS_i	The pseudonym generated for a vehicle i by RA
$Cert_i$	A certificate of pseudonym and public key for entity i
T_{expiry}	A lifetime for a certificate of entity i
$T_{current}$	The current time of entity i
$Sign_i(M)$	A signature for message M by entity i

Firstly, RA sets up the required parameters as follows:

- G cyclic additive group, G_T cyclic multiplicative group are both of order q, P is the generator of G and $\hat{e} : G * G \rightarrow G_T$ is an admissible bilinear map.
- RA chooses a random number s as its private key, $s \in z_q^*$ and its public key is $PK_{RA}=sP$. Then, tuple $\{q,P,G,G_T,PK_{RA},H_1,H_2,h\}$ is public parameters.

In our scheme, vehicles should be registered to the RA before joining VANET the registration procedure is accomplished under an off-line, secure model. The following steps describe the registration procedures:

- RA provides the parameters $\{q,P,G,G_T,PK_{RA},H_1,H_2,h\}$ to a $vehicle_i$ which issues a registration request.

- $Vehicle_i$ chooses a random number x as its private key $Sk_i=x$, $x \in z_q^*$ and generates its public key $Pk_i=xP$. Then, $vehicle_i$ sends its real identification, public key (RID_i, PK_i) to RA for its corresponding pseudonym and certificate.
- According to the (RID_i, PK_i), RA generates the pseudonym PS_i of $vehicle_i$ by $PS_i = RID_i \oplus H_1(sPK_i)$ and issues a certificate $Cert_i$ with a expiry date T_{expiry} to (RID_i, PK_i) by $\begin{cases} CT_i = H_1(PS_i \| PK_i \| T_{\exp iry}) \\ \sigma_i = sCT_i \\ Cert_i = (PS_i, PK_i, \sigma_i) \end{cases}$ And then RA sends the $(Cert_i, T_{expiry})$ to the requested $vehicle_i$.
- $Vehicle_i$ verifies and accepts the certificate, if $\hat{e}(\sigma_i, P) = \hat{e}(CT_i, PK_{RA})$, since $\hat{e}(\sigma_i, P) = \hat{e}(sCT_i, P) = \hat{e}(CT_i, P)^s = \hat{e}(CT_i, sP) = \hat{e}(CT_i, PK_{RA})$.

In order to improve the convenience and efficiency in certificate update, vehicles regenerate public/private keys and renew the pseudonym from RA according to the period of lifetime T_{expiry} which can be assigned as a specified lifetime.

- $Vehicle_i$ chooses $x_{i,j}$ as its j-th private key $Sk_{i,j}=x_{i,j}$, $x_{i,j} \in z_q^*$ and generates the public key $Pk_{i,j}=x_{i,j}P$. Then, $vehicle_i$ sends the $(Cert_{i,j-1}, T_{expiry,j-1})$ to RA.
- RA checks the current time, $Cert_{i,j-1}$ and obtains RID_i from $RID_i = PS_{i,j-1} \oplus H_1(sPK_{i,j-1})$. Then, RA generates the j-th pseudonym $Ps_{i,j}$ for $vehicle_i$ as registration procedures.

When a $vehicle_i$ want to communicate with other vehicle or RSUs, it transmits the message M, its $Cert_i$ and the signature of M as $(M, T_{current}, Cert_i, T_{\exp iry(i)}, Sign_i(M))$ where $Sign_i(M) = SK_i h(M) = x_i h(M)$. The receiving vehicles or RSU authenticate $vehicle_i$ by inspecting $\hat{e}(\sigma_i, P) = \hat{e}(CT_i, PK_{RA})$. If true, they continue to check the signature of M whether $\hat{e}(Sign_i(M), p) = \hat{e}(h(M), PK_i)$ as the following equation $\hat{e}(Sign_i(M), p) = \hat{e}(x_i h(M), P) = \hat{e}(h(M), P)^{x_i} = \hat{e}(h(M), x_i P) = \hat{e}(h(M), PK_i)$ and if it is true, they accept the message.

It is noted that the real ID of $vehicle_i$ is hided under the pseudonym PS_i which protects $vehicle_i$'s privacy. However, if necessary, other vehicles or RSU can provide the $Cert_i$ to RA to request the real ID of $vehicle_i$ which RA can restore by $RID_i = PS_i \oplus H_1(sPK_i)$.

4 Analysis and Evaluation

In our proposed scheme, RA acts as the trusted third party. According RA's public key, vehicles authenticate each other by verifying their certificates. RA uses its private key to generate vehicle's pseudonym and issues certificate within an expiry date. In the one hand, the adversaries are impossible to get the vehicle's real identity through pseudonym because of they do not hold the RA's private key. In the other

hand, when some situations need to trace vehicle's real identity, RA can reveal the real identity from pseudonym through RA's private key.

Adversaries also could not abuse the vehicle's pseudonym to send false messages, because they do not know the private key of the legal vehicle to create true *Sign(M)*. Besides, if an adversary try to update vehicle's public key and pseudonym by eavesdropping previous certificate. The RA can perceive that the certificate update request is illegal because of the real identity and private key of legal vehicle never be revealed in communications. This shows that our scheme not only provides anonymous communication but also guarantees strong non-repudiation with messages.

The efficiency of proposed scheme evaluated in the message loss ratio by the ns-2 simulator. We simulated a traffic scenario on two six-lane 1500m straight roads which set a crossroad at the middle. Vehicles are randomly deployed at the roads and an RSU is located at the intersection then every 500m along each road allocate an RSU and up to 150 vehicles can associate with a RSU. Each vehicle is driving speed in a range from 60 km/hr to 120 km/hr. The communication range for vehicle-to-vehicle and RSU-to-vehicle is 300m and 600m, respectively, which send message every 300ms. The channel bandwidth is 6Mb/s.

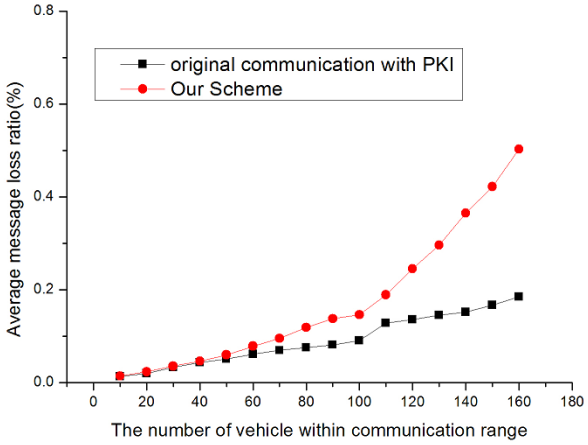


Fig. 1. Average message loss ratio versus vehicle number

Message Loss Ratio: The metric of average message loss ratio (AMLR) is defined as follows $AMLR = \frac{1}{N}\sum_{i=1}^{N}(\frac{M_{ap}^{i}}{M_{send}^{i}})$ where N is the total number of vehicles in this simulation, M_{send}^{i} represents the total number of messages have been sent by the $vehicle_i$ from the medium access control layer, the total number of messages processed by the $vehicle_i$ in application layer is denoted as M_{ap}^{i}. Figure 1 shows the relationship between the average message loss ratio. The difference between both communication model is not conspicuous until the vehicles increase to 100, the difference becomes apparent and the difference reaches about 30% when the number of vehicles up to 160. However, the normal traffic load is below 50 vehicles, where the loss ratio is merely achieved 8%.

5 Conclusion and the Future Works

In this paper we have proposed an anonymous communication scheme with non-Reputation for VANETs. One of the significant advantages is the privacy preservation that RA uses its private key to issue vehicle's pseudonym and certificate. Another is the non-repudiation; vehicles authenticate each other by verifying their certificates and RA's public key. As a future work, we plan to investigate further on how this scheme can be adapted to the message aggregation for group communications in VANETs, which is required in high density vehicle-to-vehicle communication scenarios.

References

1. Lin, X., Lu, R., Zhang, C., Zhu, H., Ho, P.H., Shen, X.S.: Security in Vehicular Ad Hoc Networks. IEEE Communications Magazine 46(4), 88–95 (2008)
2. Yang, H., Luo, H., Ye, F., Lu, S., Zhang, L.: Security in Mobile Ad Hoc Networks: Challenges and Solutions. IEEE Wireless Communications 11(1), 38–47 (2004)
3. Liu, B., Khorashadi, B., Ghosal, D., Chuah, C.-N., Zhang, M.H.: Assessing the VANET's Local Information Storage Capability under Different Traffic Mobility. In: INFOCOM 2010, pp. 1–5 (2010)
4. Raya, M., Papadimitratos, P., Hubaux, J.P.: Securing Vehicular Communications. IEEE Wireless Communications 13(5), 8–15 (2006)
5. Papadimitratos, P., Kung, A., Kargl, F., Ma, Z., Raya, M., Freudiger, J., Schoch, E., Holczer, T., Buttyan, L., Hubaux, J.P.: Secure Vehicular Communication Systems: Design and Architecture. IEEE Communications Magazine 46(11), 100–109 (2008)
6. Yeh, C.-H., Huang, Y.-M., Wang, T.-I., Chen, H.: A secure Wireless Communication Scheme for Vehicle Ad Hoc Networking. Mobile Networks and Applications 14(5), 611–624 (2009)
7. Yeh, C.-H., Hsieh, M.-Y., Li, K.-C.: A Certificate Enhanced Group Key Framework for Vehicular Ad Hoc Networks. In: Han, Y.-H., Park, D.-S., Jia, W., Yeo, S.-S. (eds.) Ubiquitous Information Technologies and Applications. LNEE, vol. 214, pp. 215–222. Springer, Heidelberg (2013)
8. Xi, Y., Sha, K., Shi, W., Schwiebert, L., Zhang, T.: Enforcing Privacy Using Symmetric Random Key-Set in Vehicular Networks. In: Proc. of the 8th International Symposium on Autonomous Decentralized Systems (ISADS 2007), Sedona, AZ, USA, pp. 344–351 (March 2007)
9. Zhang, C., Lu, R., Ho, P.H., Chen, A.: A Location Privacy Preserving Authentication Scheme in Vehicular Networks. In: Proc. of WCNC 2008, pp. 2543–2548 (2008)
10. Raya, M., Hubaux, J.-P.: Securing vehicular ad hoc networks. Journal of Computer Security, Special Issue on Security of Ad Hoc and Sensor Networks 15(1), 39–68 (2007)
11. Lin, X., Sun, X., Ho, P.H., Shen, X.: GSIS: A Secure and Privacy-Preserving Protocol for Vehicular Communications. IEEE Transactions on Vehicular Technology 56(6), 3442–3456 (2007)
12. Li, C.T., Hwang, M.S., Chu, Y.P.: A secure and efficient communication scheme with authenticated key establishment and privacy preserving for vehicular ad hoc networks. Comput. Commun. 31(12), 2803–2814 (2008)

A Mobility Management Scheme for Internet of Things

Yuan-Kai Hsiao and Yen-Wen Lin

Department of Computer Science, National Taichung University of Education, Taichung, Taiwan, R.O.C.

Abstract. In this paper, a new mobility management scheme is proposed for IoT (Internet of Things) environment. Usually, the mobile nodes are resource-limited. Specially, the mobile nodes may move together with human or vehicles in IoT. To provide ubiquitous IoT services, a network-based mobility management scheme supporting global mobility and group mobility is proposed in this paper. The performance analysis shows that the proposed scheme improves handover delay and control overhead in the context of IoT.

Keywords: IoT, Network Based, Global Mobility, Group Mobility.

1 Introduction

IoT (Internet of Things) applications [1] attract great interests recently. In IoT, things are able to communicate with each other to transparently provide ubiquitous services. IP-based networks are widely accepted for integrating existed/existing technologies to support IoT applications [2]. ITS (Intelligent Transportation System) is one of the most popular IoT applications. By integrating the technologies of IoT, vehicular networks, and cloud computing, ITS can supply real-time traffic information, monitor vehicle state, manage fleet, improve transportation efficiency, and reduce air pollution [3]. However, diverse mobility patterns in IoT impact the quality of the services. An efficient mobility management scheme is essential to support seamless services in IoT.

Mobile IP [4] is the most well-known protocol for managing mobility in IP networks. Mobile IP has been developed to meet various demands of different applications. Unfortunately, Mobile IP and its extensions are categorized as the host-based mobility protocols. That is, the mobile nodes need to be involved in the host-based mobility management processing. It seems quite unsuitable for the resource-constrained mobile nodes [5].

To remedy this problem, PMIP [6] which is a network-based mobility management protocol is proposed. In PMIP, the MAG (Mobile Access Gateway) handles handover mechanism on behalf of the mobile nodes. Consequently, the signalling overhead and energy consumption on the mobile nodes are significantly reduced. However, PMIP is a local mobility protocol that limits the mobile nodes to move across different PMIP domains [7]. When PMIP domain gets large, heavy routing overhead and system load result in poor performance [7]. Therefore, local mobility protocols are not satisfactory to provide ubiquitous services in future IoT environment.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_85,

IETF proposed PMIP-MIP Interaction [8] that integrates PMIP and Mobile IP to support global mobility. However, in PMIP-MIP, the mobile nodes are responsible of resource-consuming binding updating procedure. Also, group mobility is not considered in the original PMIP-MIP design. Oppositely, NEMO [9] supports network mobility; that managing the mobility of an entire network moving as a unit. The mobile network is connected to the Internet via the Mobile Routers (MR) being responsible of maintaining the Internet connectivity for the entire mobile network [9].

In this paper, a new mobility management scheme is proposed for IoT/ITS environment. By integrating the PMIP-MIP [8] and NEMO [9], the proposed scheme supports the aspects, including network-based, group mobility, and global mobility, for managing complicated mobility in the context of IoT/ITS. As will be shown in the performance analysis, as compared with the PMIP-MIP, the proposed scheme can shorten handover delay and lessen control overhead.

The remainder of this paper is organized as follows. In section 2, the scheme proposed in this paper is overviewed. Performance analysis is offered in section 3. A brief conclusion is presented in section 4.

2 System Overview

2.1 System Model

Fig. 1 shows the typical usage scenario of this paper. In the model, the users take the high speed train for transportation. Each user is equipped with a few bio-sensors for detecting the physiological signals of the user body (e.g. body temperature, heartbeat etc.) and a smart device (e.g. smart phones, smart watches etc.) These bio-sensors and the smart device are organized as a PAN (Personal Area Network). The smart device acts as the sink node of the PAN; which collects the signals detected by the bio-sensors and forwards them to the Internet. Numerous WiFi/WiMAX APs (Access Points) or Cellular BSs (Base Stations) are deployed alongside the road for receiving/sending the data from/to the smart device to/from the Internet. Note that, in this scenario, these bio-sensors move together with the user body. And, the users move together with the train. The train moves through a few roadside APs/BSs. Diverse movement patterns are implied in this scenario and are considered in this paper.

2.2 Network Architecture

The network architecture of the proposed scheme integrates PMIP-MIP [8] and NEMO [9] to support various movements in the context of IoT/ITS. The system is composed of a few PMIPv6 domains. *PMIPv6 domain* here is defined as the administrative area of a LMA (Local Mobility Anchor). That is, each PMIPv6 domain [6, 7] is managed by a LMA. One or more MAGs (Mobile Access Gateways) are contained in a domain. In the proposed system, the roadside APs/BSs act as the MAG.

Note that, a train contains a few carriages. The users equipped with bio-sensors are distributed in these train carriages. The communication environment of the train, as

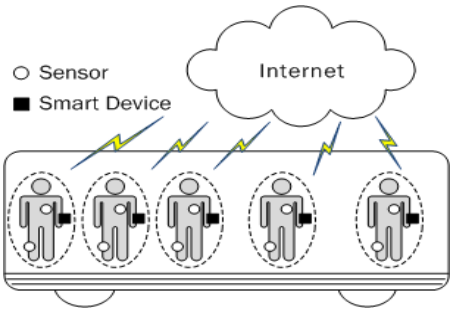


Fig. 1. System Model

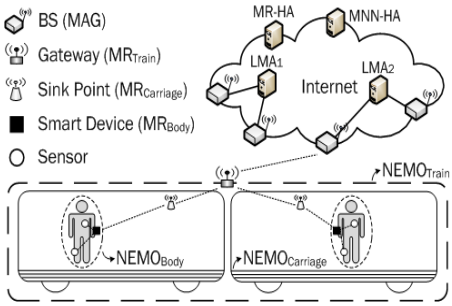


Fig. 2. Network Architecture

depicted in Fig. 2, is based on a three-layer NEMO network. Firstly, the $NEMO_{Body}$ is the PAN which is formed by the group of the bio-sensors on a user body. Each $NEMO_{Body}$ is managed by an MR_{Body}. In our design, the smart device carried with the user acts as the MR_{Body} of the $NEMO_{Body}$. The MR_{Body} relays messages between the $NEMO_{Body}$ and the $MR_{Carriage}$. Secondly, all the $NEMO_{Body}$ in a train carriage form a $NEMO_{Carriage}$. Each $NEMO_{Carriage}$ is managed by an $MR_{Carriage}$. In our design, a WiFi AP is setup for acting as the $MR_{Carriage}$. The $MR_{Carriage}$ relays messages between the $NEMO_{Carriage}$ and the MR_{Train}. Thirdly, all the $NEMO_{Carriage}$ in the train form a $NEMO_{Train}$. Each $NEMO_{Train}$ is managed by an MR_{Train}. In our design, a gateway is taken as the MR_{Train}. The MR_{Train} relays messages between the $NEMO_{Train}$ and the Internet.

Particularly, the MNNs may come from different home networks. With original NEMO [9], when the train moves, the BU messages for location updating need to be sent to different MNN-HAs (Home Agent for Mobile Node) for each MNNs; that yields considerable control messages. To remedy this problem, the MR-HA (Home Agent for Mobile Router) is coined in this paper. In our design, the information of associated LMA of each MR is maintained at the MR-HA. When the train moves across different domains, the LMA sends a BU message for each MRs to related MR-HA. Therefore, the control overhead for location updating is remarkably reduced.

2.3 The Proposed Mobility Scheme

Two types of mobility, including intra-LMA mobility and inter-LMA mobility, are considered in this paper. The train moves from the previous MAG (denoted as pMAG) to the new MAG (denoted as nMAG). The movement is categorized as the *intra-LMA mobility* in case that the pMAG and the nMAG belong to the same domain (i.e. the area managed by the same LMA). Otherwise, the movement is recognized as the *inter-LMA mobility*.

Intra-LMA Mobility: The procedure handling the intra-LMA mobility is described as follows. (1) MNNs, MR_{Body}, and $MR_{Carriage}$ attach to MR_{Body}, $MR_{Carriage}$, and MR_{Train} respectively (2) MR_{Train} attaches to pMAG, (3) MR_{Train} de-attaches from pMAG when the train moves out of the communication range of pMAG, (4) pMAG sends PBU

(Proxy Binding Update) messages to LMA for updating binding cache (i.e. the PBU message notifies LMA that MNNs de-attach from pMAG), (5) LMA sends back PBA (Proxy Binding Acknowledgement) message, (6) MR_{Train} attaches to nMAG when the train moves into the communication range of the nMAG, (7) MR_{Train} sends Rtr. Sol (Router Solicitation) messages to the nMAG, (8) nMAG sends PBU messages to LMA for update binding cache (i.e. the PBU message notifies LMA that MR_{Train} attaches to the nMAG), (9) LMA sends back PBA message, (10) nMAG sends back Rtr. Adv (Router Advertisement) to the MR_{Train}.

Inter-LMA Mobility: The procedure managing the inter-LMA mobility is introduced next. (1) MNNs, MR_{Body}, and $MR_{Carriage}$ attach to MR_{Body}, $MR_{Carriage}$, and MR_{Train} respectively (2) MR_{Train} attaches to pMAG, (3) MR_{Train} de-attaches from pMAG when the train moves out of the communication range of pMAG, (4) pMAG sends PBU messages to pLMA for updating binding cache (i.e. the PBU message notifies pLMA that MNNs de-attach from the pMAG), (5) pLMA sends back PBA message, (6) MR_{Train} attaches to nMAG when the train moves into the communication range of nMAG, (7) MR_{Train} sends Rtr. Sol messages to nMAG, (8) nMAG sends PBU messages to nLMA for updating binding cache (i.e. the PBU message notifies nLMA that MR_{Train} attaches to the nMAG), (9) nLMA sends BU messages to MR-HA of the MR_{Train} for updating binding cache (for the binding entry for MR_{Train} does not exist in nLMA binding cache), (10) MR-HA sends back BA messages to nLMA, (11) nLMA sends back PBA message to nMAG, (12) nMAG sends back Rtr. Adv to MR_{Train}.

3 Performance Analysis

The performance, including the handover delay and the signalling overhead, of the proposed scheme is presented next. The *handover delay* here means the time needed by the MNNs or the MRs to finish the handover procedure. And, the *control overhead* is the number of control messages needed to finish the handover procedure. The probability of the intra-LMA mobility is assumed as ρ. Without losing the generality, in this paper, the delay for one hop is 0.01 second [10]. Besides, the number of MNN is defined as the number of the sensors on a user body (denoted as k) while the number of NEMO networks means the number of $NEMO_{Body}$ which is composed of the sensors on the user body (denoted as N_k). In related experiments, five sensors are deployed on each user body. And, T is the handover delay time. Υ is the total number of control messages. In this paper, the performance of the proposed scheme is compared with the well-known PMIP-MIP [8]. Because the group mobility processing of NEMO is adopted, the handover delay and the control overhead of the proposed scheme are much less than those of the PMIP-MIP.

Handover Delay: The handover delay of the intra-LMA mobility and the inter-LMA mobility can be given by expression (1) and expression (2) respectively. Therefore, the overall handover delay of the proposed scheme is offered by expression (3).

$$T_{\mathrm{Pr}oposedScheme}^{Intra-LMA} = T_{RtrSol}^{MR\leftrightarrow nMAG} + T_{RtrAdv}^{MR\leftrightarrow nMAG} + T_{PBU}^{MAG\leftrightarrow LMA} + T_{PBA}^{MAG\leftrightarrow LMA} \quad (1)$$

$$T_{\mathrm{Pr}oposedScheme}^{Inter-LMA} = T_{RtrSol}^{MR\leftrightarrow nMAG} + T_{RtrAdv}^{MR\leftrightarrow nMAG} + T_{PBU}^{nMAG\leftrightarrow LMA} + T_{PBA}^{nMAG\leftrightarrow LMA} + T_{BU}^{LMA\leftrightarrow MR-HA} + T_{BA}^{LMA\leftrightarrow MR-HA} \quad (2)$$

$$T_{\mathrm{Pr}oposedScheme} = \rho[T_{\mathrm{Pr}oposedScheme}^{Intra-LMA}] + (1-\rho)[T_{\mathrm{Pr}oposedScheme}^{Inter-LMA}] \quad (3)$$

Control Overhead: The control overhead of the intra-LMA mobility and the inter-LMA mobility is calculated by expression (4) and expression (5) respectively. Thus, the overall control overhead of the proposed scheme is computed by expression (6).

$$\gamma_{\mathrm{Pr}oposeScheme}^{Intra-LMA} = the\,number\,of\,(\sum_{i=1}^{Nk}[RtrSol_i + RtrAdv_i + 2(PBU_i + PBA_i)]) \quad (4)$$

$$\gamma_{\mathrm{Pr}oposedScheme}^{Inter-LMA} = the\,number\,of\,(\sum_{i=1}^{Nk}[RtrSol_i + RtrAdv_i + 2(PBU_i + PBA_i)] + BU_i + BA_i) \quad (5)$$

$$\gamma_{\mathrm{Pr}oposedScheme} = the\,number\,of\,(\rho[\gamma_{\mathrm{Pr}oposedScheme}^{Intra-LMA}] + (1-\rho)[\gamma_{\mathrm{Pr}oposedScheme}^{Inter-LMA}]) \quad (6)$$

Discussion: The proposed scheme needs less handover delay (as shown in Fig. 3) and less control overhead (as shown in Fig. 5) than those of PMIP-MIP for various k and ρ. For both schemes, both the handover delay and the control overhead decrease as the ρ increases. In PMIP-MIP, both the handover delay and the control overhead increase apparently as k increases. Oppositely, with the proposed scheme, both the handover delay and the control overhead are independent of k.

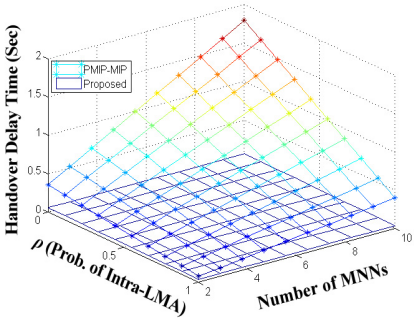


Fig. 3. Handover Delay for Various k and ρ

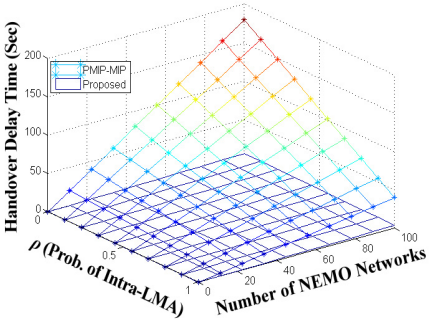


Fig. 4. Handover Delay for Various N_k and ρ

The proposed scheme needs less handover delay (as shown in Fig. 4) and less control overhead (as shown in Fig. 6) than those of PMIP-MIP for various N_k and ρ. With both schemes, both the handover delay and the control overhead decrease as the ρ increases. With the PMIP-MIP, both the handover delay and the control overhead increase remarkably as N_k increases. Relatively, in the proposed scheme, both the handover delay and the control overhead lightly increase as N_k increases.

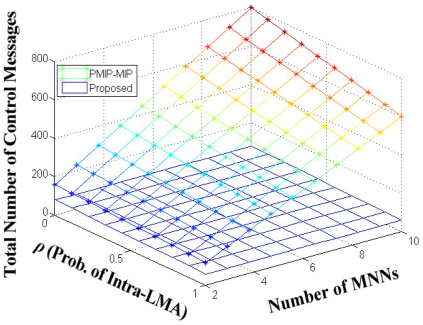


Fig. 5. Control Overhead for Various k and ρ

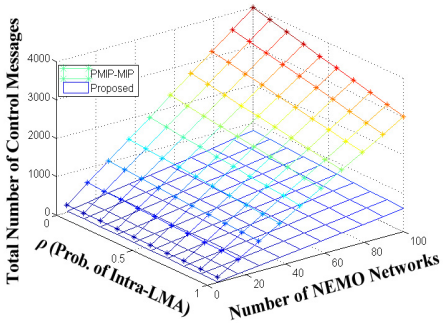


Fig. 6. Control Overhead for Various N_k and ρ

4 Conclusions

In this paper, a new network-based mobility management scheme for IoT/ITS supporting global mobility and group mobility is proposed. The performance analysis shows that the proposed scheme outperforms the PMIP-MIP in the handover delay and control overhead. The pre-handoff mechanism and the scalability problem will be considered in the future work.

Acknowledgement. This work was supported in part of by the R.O.C. National Science Council under grant number NSC 100-2221-E-142-006 and NSC 101-2221-E-142-006.

References

1. Atzori, L., Lera, A., Morabito, G.: The Internet of Things: A survey. Elsevier Computer Networks 54(15), 2787–2805 (2010)
2. Ramjee, R., La Porta, T., Salgarelli, L., Thuel, S., Varadhan, K., Li, L.: IP-based Access Network Infrastructure for Next Generation Wireless Data Networks. IEEE Personal Communications 7(4), 34–41 (2000)
3. ITS, http://en.wikipedia.org/wiki/Intelligent_transportation_system
4. Perkins, C., Johnson, D., Arkko, J.: Mobility Support in IPv6. IETF RFC6275 (2011)

5. Lee, S., Latchman, H.A., Park, B.: Efficient Handover Scheme of Proxy Mobile IPv6 in Wireless Local Area Networks. International Journal of Multimedia and Ubiquitous Engineering 5(2) (April 2010)
6. Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., Patil, B.: Proxy Mobile IPv6. IETF RFC5213 (2008)
7. Kim, H.G., Kim, J.M., Kim, H.S.: A Global Mobility Scheme for Seamless Multicasting in Proxy Mobile IPv6 Networks. In: Proceedings of the 15th International Conference on Advanced Communications Technology (2013)
8. Giaretta, G.: Interactions between Proxy Mobile IPv6 (PMIPv6) and Mobile IPv6 (MIPv6): Scenarios and Related Issues. IETF RFC6612 (2012)
9. Ernst, T., Lach, H.Y.: Network Mobility Support Terminology. IETF RFC4885 (2007)
10. Yan, Z., Zhang, S., Zhou, H., Zhang, H., You, I.: Network Mobility Support in PMIPv6 Network. In: Proceedings of the 6th International Wireless Communications and Mobile Computing Conference, pp. 890–894 (June 2010)

An Overlay Network Based on Arrangement Graph with Fault Tolerance

Ssu-Hsuan Lu[1], Kuan-Ching Li[2], Kuan-Chou Lai[3], and Yeh-Ching Chung[1]

[1] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
shlu@sslab.cs.nthu.edu.tw, ychung@cs.nthu.edu.tw
[2] Department of Computer Science and Information Engineering,
Providence University, Taichung, Taiwan
kuancli@pu.edu.tw
[3] Department of Computer Science, National Taichung University, Taichung, Taiwan
kclai@ntcu.edu.tw

Abstract. As people change the habit of using the Internet, network technology has become matured. Unlike client-server, peer-to-peer (P2P) technology increases the convenience of people's daily life. The routing efficiency of P2P system without centralized server always is an important issue. This paper proposes a virtual peer mechanism of P2P overlay network based on the arrangement graph to make exiting physical peers be agent peers for vacant peers. Each vacant peer is managed by a physical peer who often is its neighbor, and the vacant peer is called *virtual peer*. Physical peers and virtual peers make the arrangement graph full, and make the number of routing hops can be limited within the diameter of the arrangement graph. From experimental results, this system can keep routing efficiency no matter the number of peers and do not increase system overhead.

Keywords: Peer-to-Peer, overlay network, arrangement graph, fault tolerance.

1 Introduction

With the emergence of the Internet, the traditional lifestyle has been subverted, and people have been more and more depended on the Internet. Due to the efficiency driven, people gradually abandon the use of client-server architecture, and use peer-to-peer (P2P) architecture. In such an environment, all peers are both client and server, and then can share resources and information directly.

P2P overlay networks are virtual networks based on physical networks. Through the overlay network, distributed peers are connected and operated like being in the same domain. In P2P overlay networks, peers join or depart frequently, so fault tolerance is an important research issue.

Previous work introduces a P2P overlay network which is developed based on arrangement graph [2], inherits properties of the arrangement graph, and provides great performance [7]. Therefore, this study utilizes a dynamical scaling mechanism and the concept of virtual peers to deal with vacant peers to make the arrangement graph can be kept full for achieving the effect of fault tolerance.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_86,

According to the rule of the arrangement graph, in this system, each peer has a unique ID for identification, and there are two important parameters of the arrangement graph, n and k. The parameter k indicates the number of digits of each peer ID, and n is the range of each digit, with $1 \leq k \leq n$-1. Only one digit of the peer IDs of any two adjacent peers is different. Furthermore, the maximum number of peers that the system can accommodate is $\frac{n!}{(n-k)!}$, the degree of each peer is $k(n\text{-}k)$, and the diameter which means the longest distance between any two peers is equal to $\left\lfloor \frac{3k}{2} \right\rfloor$.

The dynamical scaling mechanism can make our system adjust the values of its parameters according to the number of peers in the system. Furthermore, the concept of virtual peers is used in this study to make existing physical peers act as agents of vacant peers. Those peers who are managed by other physical peers are called *virtual peers*. After a new peer joins the system, it tries to discover information of its neighbors, and becomes agents of its vacant neighbors if it is needed. By using this method, the arrangement graph can be kept full, and the routing hops of the arrangement graph can be kept within $\left\lfloor \frac{3k}{2} \right\rfloor$. When a peer leaves the system, its neighbors will compare their statuses and decide which neighbor will become the agent of that peer. The peer who acts less virtual peers gains the right of agent.

The remainder of this paper is organized as follows. Section 2 presents details on P2P overlay networks and the arrangement graph. Section 3 describes the proposed method, and some simulation results are shown in section 4. Finally, conclusions are discussed in section 5.

2 Related Work

In this section, some introductions of P2P overlay networks and the arrangement graph are introduced. According to the topologies of P2P overlay networks, P2P overlay networks can be classified as *structured* and *unstructured* overlay networks, and each of them has a specific method for maintaining and routing.

In the structured overlay network, the locations of contents/files are tightly controlled by using specific mapping method. The most common method is to use Distributed Hash Table (DHT) [6] to map contents/files and peers. Chord [9], Pastry [8], and Koorde [5] are notable examples of structured P2P overlay networks. In the unstructured P2P overlay network [3, 4], there is no such relationships between contents/files and peers. Peers only know information of their neighbors, so peers often use flooding to transmit queries on overlay networks, such as gnutella [6].

The arrangement graph is denoted by $A_{n,k}$ which was proposed by Day and Tripathi as a generalization of the star graph. Let n and k be two positive integers with $k \leq n$-1, and let sets $\langle n \rangle$ and $\langle k \rangle$ denote the sets $\{1, 2, ..., n\}$ and $\{1, 2, ..., k\}$, respectively.

An (n, k)-arrangement graph, $A_{n,k}$, is an undirected graph (V, E), where $V = \{p_1 p_2 ... p_k \mid p_i \in \langle n \rangle$ and $p_i \neq p_j$ for $i \neq j\} = P_k^n$, and $E = \{(p, q) \mid p, q \in V$, and for some i in $\langle k \rangle$, $p_i \neq q_i$ and $p_j = q_j$ for $j \neq i\}$. Let P_k^n be the set of permutations of k elements taken from $\langle n \rangle$.

G=(V, E) is the set of peers, such as participating peers, and edges, such as overlay links. Each peer of $A_{n,k}$ is an arrangement with k digits out of n elements of $\langle n \rangle$, and the edges connect peers with only one different digit. $A_{4,2}$-arrangement graph is shown in Fig. 1. The degree of each peer is four, so each peer has four neighbors. The peer 24 is connected to the peers 14, 23, 21, and 34. For peer 24, there is only one digit different from the four peers 14, 23, 21, and 34. In other words, all peers that have one different digit are connected.

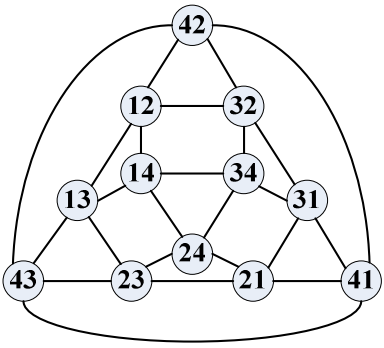


Fig. 1. $A_{4,2}$-arrangement graph

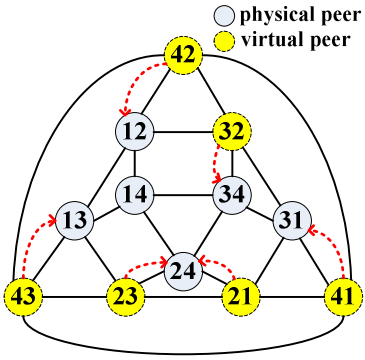


Fig. 2. Illustration of the Virtual Peer Mechanism

3 Virtual Peer Mechanism

The dynamical scaling mechanism is utilized to adjust the system scale when the number of peers achieves a certain extent. When the number of peers achieves 90% of the maximum number of peers that the system can accommodate, the system extends its system scale. Similarly, when the number of peers only remains 20% of the maximum number of peers that the system can accommodate, the system shrinks its system scale. Furthermore, in this study, the concept of virtual peers is used in our system. The virtual peers mean peers that do not really exist in the system and are managed by other physical peers. In this way, physical peers and virtual peers can make the arrangement graph full. Furthermore, the routing efficiency can be limited within $\left\lfloor \frac{3k}{2} \right\rfloor$ hops, which is the same as the diameter of the arrangement graph. Therefore, this method can ensure data completeness and maintain the property of the arrangement graph to achieve fault tolerance. The following describe some actions for maintaining virtual peers.

When a new peer joins the system, it runs through a series of actions to obtain its peer ID, and generates its neighbor table according to its peer ID. After that, the new peer also needs to discover information of its neighbors. Some neighbor peers already exist in the system, so the new peer records their information. But some other neighbor peers do not exist, and then the new peer receives the information of agents of those neighbor peers who do not exist. However, if the new peer finds that the

agent of a neighbor peer is not neighboring relation, the new peer tries to obtain the right of managing that neighbor peer.

On the other hand, before a peer leaves the system, it chooses a physical neighbor peer to be the agent for it. If all of its neighbor peers are virtual peers, it also needs to choose a neighbor to become the agent for it. Besides, the leaving peer also needs to choose some other peers to be agents for virtual peers it manages.

But, sometimes a peer leaves the system in abnormal ways, and is unable to do the above actions before it leave. Its neighbors still can fix this problem, because peers send messages to their neighbors periodically to check whether their neighbors still exist. All neighbor peers of the leaving peer send messages to each other, and decide who can be the agent of the leaving peer. Physical neighbor peers can have the highest priority, and the number of virtual peers they already manage also be considered. After deciding the new agent, the new agent sends messages to the other neighbors of the leaving peer to inform them of the newest information.

Fig. 2 shows the concept of the virtual peer mechanism. Peers 12, 13, 14, 24, 34, and 31 are physical peers that already exist in the system, and other peers are virtual peers that are managed by physical peers. For example, peer 23 is a virtual peer that is managed by peer 24, and peer 24 is the agents of peer 23 and 21.

4 Experimental Results

In this section, some experimental results are presented to demonstrate the benefits of the proposed system. OverSim [1, 10] was used to evaluate the performance of the proposed method and other systems, because it allows them to be evaluated based on the same environment setup.

4.1 Experimental Environment

OverSim is an open-source simulation framework for building overlay and P2P network on top of the OMNeT++ simulation environment [11], and is a powerful and widely used simulator for investigations of P2P environments because it contains several models for both structured and unstructured P2P systems. Our proposed method was compared with the AGO [7], Chord, and Koorde. Some related settings of parameters are as following: the parameter of *m* in Chord and Koorde is *14*, AGO were executed with $A_{n,k} = A_{9,5}$, and our proposed method was initiated with $A_{n,k} = A_{6,5}$ and extend to $A_{9,5}$ gradually by using dynamical scaling mechanism. These parameters are set to make the peer space of each overlay network is close to 20,000 peers.

4.2 Average Routing Hops

Efficient routing performance is very important in P2P overlay networks especially in a large-scale environment. Fig. 3 shows the average number of routing hops required

to look up information for each P2P overlay network with a different number of peers. From Fig. 3, the virtual peer mechanism indeed can eliminate the problem of detour, because the proposed method decreases the average number of routing hops of AGO about 1 hop by improving the problem of detour in AGO resulted from vacant peers. Because virtual peers can serve routing requests and guarantee the shortest path, the proposed method can also perform better than Chord and Koorde.

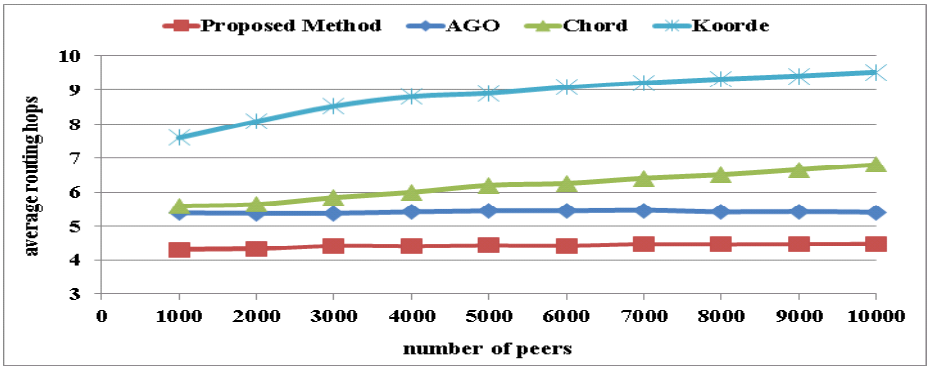


Fig. 3. Average number of routing hops required to look up information

4.3 Bandwidth Consumption

The bandwidth consumption also is an important issue in P2P overlay networks, because it often affects the performance of message transmission. The bandwidth consumption is the average sent and received message rate of a peer. In P2P environment, peers join or leave frequently and P2P systems need to maintain that information which makes extra overheads of P2P systems.

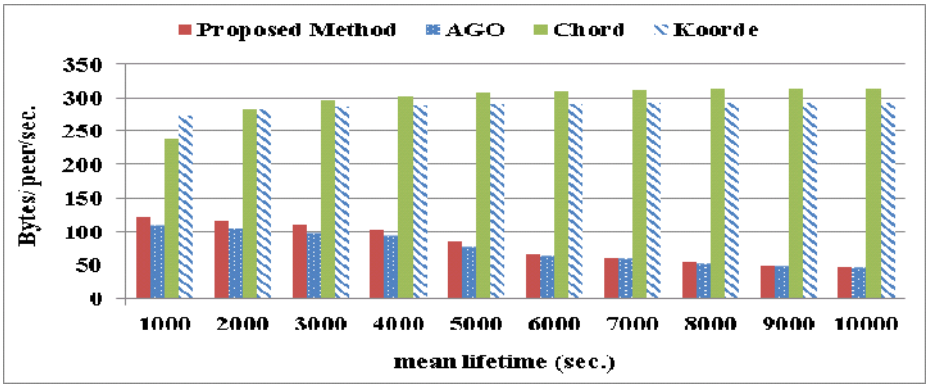


Fig. 4. Average message sizes of each peer (bytes)

Fig. 4 shows average message sizes of each peer which were run with 10,000 peers. These experiments were run with different churn rates to illustrate the influence of dynamical environment. From Fig. 4, the proposed method and AGO produce the least message sizes, and the bandwidth consumption of the proposed method is almost the same as AGO's. The proposed method can decrease message sizes due to the dynamical scaling mechanism, but increase message sizes for maintaining virtual peers. Therefore, the proposed method produces the same or a little more message sizes than AGO. But the proposed method indeed can improve routing performance of AGO.

5 Conclusions

This study proposes the virtual peer mechanism and the dynamical scaling mechanism of the AGO based on the arrangement graph. Because peers often join or depart in the P2P environment, peers in the arrangement graph may be vacant and this situation affects routing performance. Therefore, virtual peers can make the routing hops be limited within $\left\lfloor \frac{3k}{2} \right\rfloor$ hops the same as the diameter of the arrangement graph. The experimental results illustrate that these two mechanisms indeed can perform routing well even in a large-scale, dynamical environment.

References

1. Baumgart, I., Heep, B., Krause, S.: OverSim: A Flexible Overlay Network Simulation Framework. In: The Proceedings of 10th IEEE Global Internet Symposium (GI 2007) in conjunction with IEEE INFOCOM 2007, Anchorage, AK, USA, pp. 79–84 (May 2007), doi:10.1109/GI.2007.4301435
2. Day, K., Tripathi, A.: Arrangement graphs: A class of generalized star graphs. Inform. Processing Lett. 42, 235–241 (1992)
3. Haribabu, K., Reddy, D., Hota, C., Yla-Jaaski, A., Tarkoma, S.: Adaptive Lookup for Unstructured Peer-to-Peer Overlays. In: The Proceedings of 3rd International Conference on Communication Systems Software and Middleware and Workshops 2008 (COMSWARE 2008), January 6-10, pp. 776–782 (2008)
4. Jiang, S., Guo, L., Zhang, X.: LightFlood: Minimizing Redundant Messages and Maximizing Scope of Peer-to-Peer Search. The Journal of IEEE Transactions on Parallel and Distributed Systems (TPDS) 19(5), 601–614 (2008)
5. Frans Kaashoek, M., Karger, D.R.: Koorde: A simple degree-optimal distributed hash table. In: Kaashoek, M.F., Stoica, I. (eds.) IPTPS 2003. LNCS, vol. 2735, pp. 98–107. Springer, Heidelberg (2003)
6. Lua, E.K., Crowcroft, J., Pias, M., Sharma, R., Lim, S.: A Survey and Comparison of Peer-to-Peer Overlay Network Schemes. The Journal of IEEE Communications Survey and Tutorial 7(2), 72–93 (Second Quarter, 2005)
7. Lu, S.-H., Lai, K.-C., Li, K.-C., Chung, Y.-C.: Design and analysis of arrangement graph-based overlay systems for information sharing. In: The Proceedings of the 3rd IEEE International Workshop on Management of Emerging Networks and Services (IEEE MENS 2011) in conjunction with IEEE GLOBECOM 2011, Houston, Texas, USA, pp. 694–698 (2011)

8. Rowstron, A., Druschel, P.: Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. In: Guerraoui, R. (ed.) Middleware 2001. LNCS, vol. 2218, pp. 329–350. Springer, Heidelberg (2001)
9. Stoica, I., Morris, R., Liben-Nowell, D., Karger, D., Frans Kaashoek, M., Dabek, F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. The Journal of IEEE/ACM Transactions on Networking 11(1), 17–32 (2003)
10. http://www.oversim.org/
11. http://www.omnetpp.org/

Event Detection in Wireless Sensor Networks: Survey and Challenges

Aziz Nasridinov[1], Sun-Young Ihm[1], Young-Sik Jeong[2], and Young-Ho Park[1]

[1] Department of Multimedia Science, Sookmyung Women's University,
Cheongpa-ro 47-gil 100, Yongsan-Ku, Seoul, 140-742, Korea
{aziz,sunnyihm,yhpark}@sookmyung.ac.kr
[2] Department of Multimedia Engineering, Dongguk University,
30, Pildong-ro 1-gil, Jung-gu, Seoul 100-715, Korea
ysjeong2k@gmail.com

Abstract. In typical wireless sensor networks (WSNs), sensor nodes have limited resources such as battery power, computing capability and memory. Creating an event detection method comprising with those resource limitations is not an easy task and this sets several challenges. In this paper, we first describe challenges in event detection in WSNs. Then, we investigate the previous studies that have been done for solving those challenges.

Keywords: Event detection, wireless sensor networks, survey.

1 Introduction

Wireless sensor networks (WSNs) are constructed of many tiny devices, called sensor nodes, randomly distributed over a large location. The sensor nodes are equipped with a sensing, data communicating, and processing units, which enable them to monitor the physical world, communicate and exchange the collected sensory data with other nodes, locally process and make decisions about monitored phenomena. This will lead to the detection of events and unusual data behaviors in a monitored environment. This feature is referred as *event detection*. Event detection in WSNs has received much attention in variety of applications, such as military target tracking and surveillance, meteorological hazards, wildlife monitoring, natural disaster relief and healthcare.

There are following requirements that should be satisfied in the event detection: timeliness, a high true detection rate, and a low false alarm rate [1]. However, in typical WSNs, sensor nodes have limited resources such as battery power, computing capability and memory. Creating an event detection method comprising with those resource limitations is not an easy task and this sets several challenges. Kerman et al. [6] described the most common challenges in event detection as: situational dependence, criticality of application, numerous and diverse data sources, and network topology. In this paper, we give a detailed explanation to the challenges in event detection in WSNs. Then, we investigate the previous works that have been done for solving those challenges.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_87,

The rest of the paper is proceeds as follows. Chapter 2 discusses challenges in event detection in WSN. Chapter 3 presents existing solutions. Chapter 4 highlights conclusions.

2 Challenges in WSNs Event Detection

Sensor nodes possess limited resources such as battery power, computing capability and memory. Creating an event detection method dealing with with those resource limitations is not a trivial task and this sets several challenges. Kerman et al. [6] described the most common challenges in event detection as: situational dependence, criticality of application, numerous and diverse data sources, and network topology. In this section, we give a detailed explanation to those challenges (Table 1).

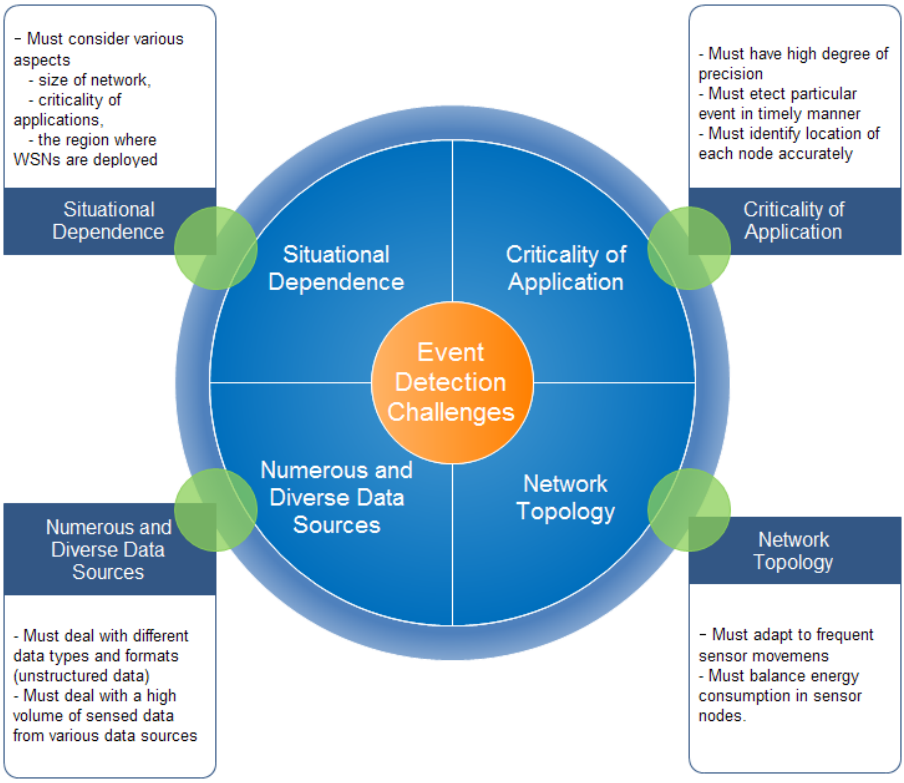


Fig. 1. Challenges in WSNs event detection [2]

Situational Dependence: Since, WSNs are used in a wide range of applications, event detection can be different according to specific situation. For instance, some metrics that are used to detect an event in military target tracking and surveillance cannot be

applied to the healthcare applications. The network administrator must consider many different aspects in such environments. Among those aspects are the size of network, criticality of applications, and the region where WSNs are deployed (e.g. WSNs can be deployed in a plain land or in mountain region). Thus, designing an event detection approach that is adaptable to various situations is a challenging task.

Criticality of Application: WSNs are utilized in many critical applications such as measuring indicators of imminent catastrophic machine failures, detecting breaches within security perimeters, and observing human stasis parameters. Event detection approach in such critical applications must have a high degree of precision and detect particular event in timely manner. Moreover, location identification of each node provides the base for routing, density control, tracking, and number many of other communication network aspects. Thus, it is important in critical applications that each node reports its location accurately.

Numerous and Diverse Data Sources: WSNs can be deployed in various locations and consist of hundreds to thousands of sensor nodes meaning that there is a high volume of sensed data from various data sources. Moreover, sensor data can be unstructured meaning that it can contain video, images, text documents, audio, multivariate records, relational data, and spatio-temporal data. Thus, event detection approach must comprise different data types and formats coming from numerous and diverse data sources.

Network Topology: In typical WSNs applications, it is very important to construct efficient network topology, because it would reduce energy consumption and prolong network lifetime. However, in some WSNs, such the network topology changes frequently due to sensor mobility and sensor lifetime. Moreover, there is heterogeneity among sensor nodes such as different residual energy, transmission speed, transmission range, and nodal traffic. Thus, in such dynamic and heterogeneities WSNs, event detection approach must act according to movement of sensor node outside of the intended observed area, power consumption, sensor failures and finite sensor lifetimes, and balance the energy consumption in sensor nodes.

3 Event Detection Methods in WSNs

Most event detection methods fit into one of following three categories: statistical, probabilistic, and artificial intelligence and machine learning. This section describes those event detection methods in details.

A. Statistical Methods

Gupchup et al. [3] proposed event detection approach based on statistical signal processing techniques. They used Principal Component Analysis (PCA) technique to build a compact model of the observed phenomena that detects a various events from the collected measurements in environmental monitoring, such as seasonal trends or rain events. The authors use the divergence between actual collected measurements and model predictions to detect the existence of discrete events within the collected

data streams. The experiment results show that proposed approach is able to detect the onset of rain events using the temperature modality of a wireless sensor network.

Meng et al. [8] proposed a compressive sensing method for sparse event detection in WSNs. In their approach, they formulate two problems of WSNs. First, there are a small number of active sensor nodes comparing to the total sensor nodes. Second, different events may occur concurrently and lead to the interference in detecting them individually. In order to solve these problems, the authors adopted a marginal likelihood maximization algorithm and a heuristic algorithm for the Bayesian framework. The authors insist that their approach lead to the higher detection probability than the traditional linear programming solution.

Vu et al. [9] study the Timely Energy-efficient k-Watching Event Detection problem (TEKWED) for composite event detection and alarming in WSNs. In order to solve this problem, they proposed a novel scheme that is able to detect events and deliver timely warnings in WSNs. Based on this scheme, an algorithm that considers topology of the network and routing capabilities is proposed. This algorithm builds a set of detection sets that have several advantages, including the short notification time, energy conservation, and tunable quality of surveillance requirements in event alarming applications.

B. Probabilistic Methods

Li et al. [7] deals with detecting of complex events and proposed a non-threshold based approach for complex event detection in 3D environment monitoring applications. The authors use energy-efficient methods to collect a time series of data maps from the sensor network and detect complex events through matching the gathered data to spatio-temporal data patterns.

Zoumboulakis et al. [10] proposed an approach for detecting complex events in sensor networks. They define complex events as sets of data points hiding interesting patterns and insist that these unusual patterns are difficult to detect using existing technologies. Inspired from time-series data mining techniques, the authors proposed an approach to convert raw real-valued sensory data to symbolic representations using a Symbolic Aggregate Approximation (SAX) algorithm and then use a distance metric for string comparison.

Ihler et al. [4] proposed a framework for unsupervised learning in this context, based on a time-varying Poisson process model that can also account for anomalous events. The authors demonstrate how the parameters of this model can be learned using statistical techniques. Through the extensive experiments, the authors show that proposed approach performs significantly better than a non-probabilistic and threshold-based technique. Moreover, the proposed model can be utilized in examining different degrees of periodicity in the data, and make inferences about the detected events.

C. Artificial Intelligence and Machine Learning Methods

Bahrepour et al. [2] studied the role of machine learning techniques in event detection and proposed an approach for detecting disastrous using WSNs. Specifically,

proposed approach is based on detecting events using decision tree classifiers running on individual sensor nodes and applying a voting to reach a consensus among detections made by various sensor nodes. The authors explain the motivation behind choosing decision trees is their simplicity and explicit form of expression as if-then-else rules that fulfill the requirements posed by resource limitations of WSNs.

Abadi et al. [1] proposed REED, a system for robust, efficient filtering and event detection in sensor networks. Their approach extends the TinyDB query processor with facilities for efficiently executing multi-predicate filtration queries inside a sensor network. The proposed approach have three main features: running in limited amounts of RAM, can distribute the storage burden over groups of nodes, and are tolerant to dropped packets and node failures. It makes it suitable a wide range of event-detection applications that traditional sensor network database systems cannot be used to implement.

Kapitanova et al. [5] described disadvantage of current event detection approaches in relying on the usage of precise values to specify event thresholds. The authors insist that t crisp values cannot adequately handle the often imprecise sensor readings. Thus, instead of crisp values, the authors propose to use a fuzzy value, which decrease the number of false positives and improves the accuracy of event detection. The authors also demonstrated the advantage of proposed approach over well-established classification algorithms.

4 Conclusion

Due to several restrictions in WSN, creating an event detection method comprising with these resource limitations is not an easy task and this sets several challenges. In this paper, we first described those challenges in event detection in WSN. Then we investigated the previous works that have been done for solving these challenges and describe their limitations.

Acknowledgement. This work was supported by the IT R&D program of MKE/KEIT. [10041854, Development of a smart home service platform with real-time danger prediction and prevention for safety residential environments].

References

1. Adabi, D., Madden, S., Lindner, W.: REED: Robust, Efficient Filtering and Event Detection in Sensor Networks. In: Proceedings of the 31st Very Large Databases Conference, pp. 769–780 (2008)
2. Bahrepour, M., Meratnia, N., Poel, M., Taghikhaki, Z., Havinga, P.J.M.: Distributed Event Detection in Wireless Sensor Networks for Disaster Management. In: Proceeding of the IEEE International Conference on Intelligent Networking and Collaborative Systems, pp. 507–512 (2010)

3. Gupchup, J., Burns, R., Terzis, A., Szalay, A.: Model-Based Event Detection in Wireless Sensor Networks. In: Proceedings of the Workshop on Data Sharing and Interoperability on the World-Wide Sensor Web (2008)
4. Ihler, A., Hutchins, J., Smyth, P.: Adaptive Event Detection with Time-Varying Poisson Processes. In: The Twelfth International Conference on Knowledge Discovery and Data Mining, pp. 207–216 (2008)
5. Kapitanova, K., Son, S.H., Kang, K.-D.: Event detection in wireless sensor networks – can fuzzy values be accurate? In: Zheng, J., Simplot-Ryl, D., Leung, V.C.M. (eds.) ADHOCNETS 2010. LNICST, vol. 49, pp. 168–184. Springer, Heidelberg (2010)
6. Kerman, M.C., Jiang, W., Blumberg, A.F., Buttrey, S.E.: Event Detection Challenges, Methods, and Applications in Natural and Artificial Systems. In: Proceeding of the International Command and Control Research and Technology Symposium, pp. 1–19 (2009)
7. Li, M., Liu, Y., Chen, L.: Non-Threshold based Event Detection for 3D Environment Monitoring in Sensor Networks. In: Proceedings of the 27th International Conference on Distributed Computing Systems, pp. 1–9 (2007)
8. Meng, J., Li, H., Han, Z.: Sparse event detection in wireless sensor networks using compressive sensing. In: Proceeding of Conference on Information Sciences and Systems, pp. 181–185 (2009)
9. Vu, C.T., Beyah, R.A., Li, Y.: Composite Event Detection in Wireless Sensor Networks. In: Proceeding of the IEEE International Performance, Computing, and Communications Conference, pp. 264–271 (2007)
10. Zoumboulakis, M., Roussos, G.: *Escalation:* Complex event detection in wireless sensor networks. In: Kortuem, G., Finney, J., Lea, R., Sundramoorthy, V. (eds.) EuroSSC 2007. LNCS, vol. 4793, pp. 270–285. Springer, Heidelberg (2007)

Accelerating Adaptive Forward Error Correction Using Graphics Processing Units

Md Shohidul Islam and Jong-Myon Kim*

School of Electrical Engineering, University of Ulsan,
93 Daehak –ro, Nam-gu, Ulsan 680-749, Korea
shohid@mail.ulsan.ac.kr, jmkim07@ulsan.ac.kr

Abstract. The demand of error free high-speed, real-time wireless communication is mounting day by day. Adaptive forward error correction (AFEC) is one of the error control mechanisms in which corrupted packets are automatically corrected at the destination end. Graphics processing units (GPUs) offer highly parallel computing platform, and we propose a GPU based AFEC approach for fast error recovery in this paper. We develop a massively parallel AFEC algorithm using the GPU and accomplish performance comparison with an equivalent serial algorithm that runs on the traditional CPU. Experimental results demonstrate that the proposed GPU based AFEC approach enormously outperforms the sequential approach yielding significant reduction in execution time while improving buffer utilization. In addition, the proposed GPU based approach achieves the average speedup of 74× over the sequential algorithm using the CPU while reducing the computational complexity from $O(n^3)$ of the sequential algorithm to $O(n)$ by using the single instruction multiple data (SIMD) based GPU.

Keywords: High-speed real-time wireless communication, packet corruption, AFEC, Hamming code, GPU.

1 Introduction

Wireless network is an overwhelming reality and drawing wide attention in the present day world. It ranges from wireless cellular, real time audio and video communication, mobile ad-hoc network (MANET), wireless sensor network (WSN), etc. [1]. However, due to the absence of physical media it faces some challenging issues. Signals are easily corrupted by disturbance such as interference, reflection, diffraction, multi-path fading and different noises. To overcome the adverse impact on wireless communication, error control is required [2]. More importantly, in real time and high speed networks, satellite and space communication where massive data need to be processed momentarily, it requires faster error recovery.

GPUs offer massively parallel computing platform [5], [6], [7]. Therefore, this paper proposes Hamming coding [3], [4] based parallel AFEC approach by using

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_88,

GPU. We validate our proposed approach using a compute unified device architecture (CUDA) [6], [7] enabled NVIDIA GeForce GTX 560 graphics card. We evaluate impacts of the proposed parallel AFEC approach on network performance in terms of execution time, computational complexity and speedup. Experimental results indicate that the proposed GPU based AFEC enormously outperforms the sequential approach with respect to all the considered performance metrics.

The remainder of this paper is organized as follows. Section 2 discusses the proposed approach, and Section 3 presents a brief overview of the AFEC and its implementation using CPU and GPU, and Section 4 presents experiment results and analysis. Finally, Section 5 concludes the paper.

2 The Proposed Approach

In the proposed GPU based error recovery model as shown in Fig. 1, the sender is responsible for encoding the original data with some redundant information, and the encoded information propagates through the wireless medium to the receiver. The receiver executes Hamming coding based AFEC algorithm upon the received packets using GPU for faster error detection and correction. Finally, the algorithm removes redundant information, thereby retrieving the original message. Checksum calculations, the crucial part of Hamming algorithm, spend much time on CPU than on GPU.

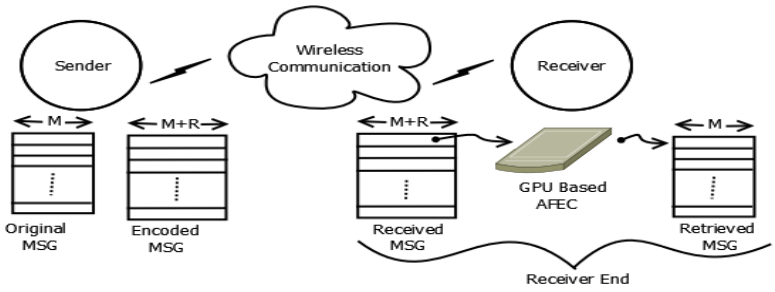


Fig. 1. The proposed GPU based network approach

3 AFEC Overview and Implementation

In this study, Hamming code is used because of its lower-order space complexity and its robustness in action [4]. This section briefly introduces the fundamental overview of Hamming code [3], [4]. The algorithm on the sender side encodes a packet of M bit using Hamming coding technique. For this purpose, R redundant bits are added with the original message to obtain Hamming coded packet, H, where the number of extra bit required is calculated from the following inequality [3].

$$2^R \geq M + R + 1 \tag{1}$$

Fig. 2 shows a typical example of sender encoding for M= 7. In this case, four redundant bits are needed namely, r_1, r_2,…,r_R. The position of the redundant bits in the hamming coded packet are determined by $r_{i+1}=2^i$ where i=0, 1, 2,…,(R-1). To calculate the value of redundant bits, all the numbers from '1' to 'M+R' (here 1 to 11 for M=7, R=4) have to be converted into R bit binary as shown in Fig. 4. For each binary sub column from LSB to MSB, we have to consider the cell containing '1'. Then the corresponding decimal numbers indicate the indices of hamming coded message. Data in the indices selected for X-OR (exclusive OR) operation to get 'r' values are shown in equation (2), (3), (4) and (5).

$$r_1 = H_3 \oplus H_5 \oplus H_7 \oplus H_9 \oplus H_{11} \tag{2}$$

$$r_2 = H_3 \oplus H_6 \oplus H_7 \oplus H_{10} \oplus H_{11} \tag{3}$$

$$r_3 = H_5 \oplus H_6 \oplus H_7 \tag{4}$$

$$r_4 = H_9 \oplus H_{10} \oplus H_{11} \tag{5}$$

Fig.2 and Fig.3 shows the packet encoding stage at the sender end. At the receiving end, packet decoding, error detection and correction are performed on CPU and GPU, respectively.

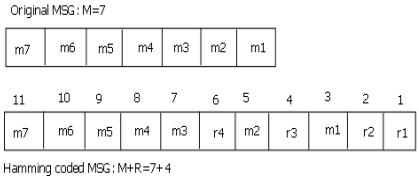


Fig. 2. Message encoding at the sender end

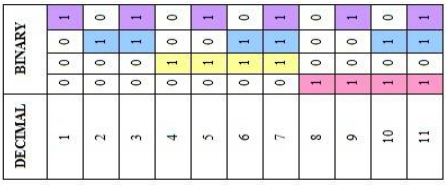


Fig. 3. Cell selection to calculate redundant bits

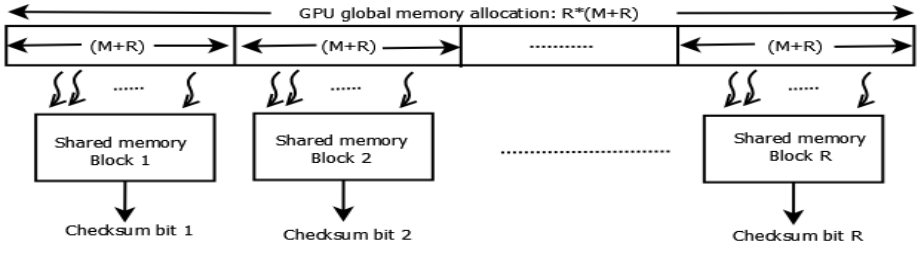


Fig. 4. Parallel AFEC mapping on GPU

GPU is virtually decomposed into many grids; grids are further partitioned into a lot of blocks; and each block contains a number of threads. A group of 32 threads

forms an execution unit called a warp, and each thread computes the same instruction synchronously [6], [7]. Each multiprocessor inside GPU can execute one or more warps concurrently. The blocks and the threads can be one-, two- or three-dimensional. Fig. 4 shows the mapping of the parallel algorithm on GPU. A total of $R \times (M + R)$ units of global memory is allocated for data transfer from CPU and R blocks are declared with dimension $M+R$, where each block is responsible to calculated one checksum bit. Checksum computation is the most task-intensive section of packet decoding and error detection. A total of R checksums is calculated by using exactly same procedure as in the sender side to generate redundant bits. GPU algorithm determines that a packet of n bits length requires $(n-1)$ computational cycles, yielding complexity of $O(n)$ for error detection and correction.

4 Experimental Results

4.1 Execution Time

Execution time is vital for time-sensitive and real-time networks. In these contexts, quality of service depends heavily on the speed of error recovery. Fig. 5 shows the execution time for different packet sizes. Execution time increases with packet size, usually for the following two reasons. Firstly, longer messages have a greater number of redundant bits attached to the message. Secondly, the code used to calculate redundant information is also lengthened, increasing the number of XOR operations required. In addition, the computational cycles are proportionally influenced by the hamming code length for any packet.

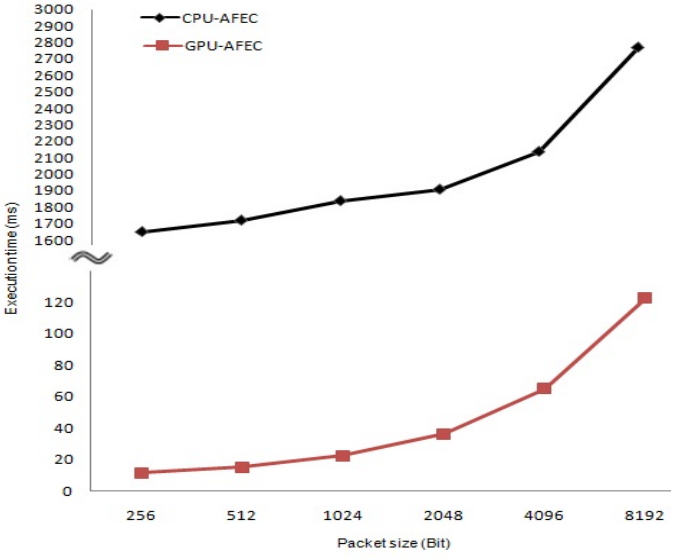


Fig. 5. Execution time of CPU and GPU

Speedup is an indication of how much faster a parallel processing is over its counterpart, i.e., sequential processing. Speedup of the GPU-based approach over the CPU-based approach is defined as ratio of the execution time of the sequential algorithm to that of the equivalent parallel algorithm. Table 1 shows the speedup obtained by the GPU algorithm over the CPU based approach for various packet size and the average speedup of the GPU algorithm achieves around 74× over the CPU approach.

Table 1. Speedup by GPU over CPU

Packet size (bit)	Speedup of GPU over CPU
256	142×
512	113×
1024	82×
2048	52×
4096	32×
8192	22×

4.2 Buffer Utilization

Buffer is another precious system resource. Upon receiving, packets are temporarily stored to the receiver buffer and wait for de-queuing, decoding, error detection, and correction. Faster error checking leads to faster de-queuing which reduces packet waiting time in the queue. High-speed, real-time data communication implies that the rate of packet incoming into the queue from medium is very high.

If χ is the total waiting time of n packets in the receiver buffer before concluding their processing in the CPU or GPU, λ_i is the waiting time of i-th packet before it is de-queued and τ_l is processing time for one packet of length l, then χ is given by

$$\chi = \sum_{\forall i \in [1,\, n]} \lambda_i = \sum_{\forall i \in [1,\, n]} (i-1)\tau_l = \frac{n(n-1)}{2} \times \tau_l \quad . \tag{6}$$

In (6), τ_l is the key determinant of χ. Packet waiting time is closely related to the buffer size. Longer buffer occupation means that its size has to be large enough to accommodate further incoming packets.

As can be seen in Table 2, τ_l for GPU is much lower than that for CPU. Table 3 shows that packets occupy the receiver buffer for very short instant of time relating to GPU processing. Advantages of waiting time minimization are many folds, such as- (a) it increases the availability of free space in buffer over time, (b) it makes the buffer even much ready to accommodate more incoming packets and (c) it enhances buffer's adaption and synchronization capability with high speed traffic flow.

Table 2. Single packet processing time

Packet length (Bit)	CPU (ms)	GPU (ms)
256	1.65	0.06
512	1.72	0.07
1024	1.84	0.07
2048	1.91	0.09
4096	2.14	0.12
8192	2.77	0.19

Table 3. Percentage of packet waiting time

Packet length (Bit)	CPU	GPU
256	100%	3.53%
512	100%	3.81%
1024	100%	3.97%
2048	100%	4.60%
4096	100%	5.45%
8192	100%	6.81%

5 Conclusions

In this paper, we proposed a novel GPU based error recovery system for real-time, high-speed wireless network, satellite and space communication, where massive data are generated and need to be processed instantaneously. We investigated the possible impacts of the parallel error correction on network performance parameters. The experimental results demonstrated that the proposed GPU based error control enormously outperforms the sequential approach using the CPU in terms of execution time, computational complexity and buffer utilization.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566), and by the Leading Industry Development for Economic Region (LeadER) grant funded by the MOTIE(The Ministry of Trade, Industry and Energy), Korea in 2013. (No. R0001220).

References

[1] Teymoori, P., Yazdani, N., Khonsari, A.: DT-MAC: An Efficient and Scalable Medium Access Control Protocol for Wireless Networks. IEEE Transactions on Wireless Communications 12(3), 1268–1278 (2013)

[2] Zou, Z., Soldati, P., Zhang, H., Johansson, M.: Energy-Efficient Deadline-Constrained Maximum Reliability Forwarding in Lossy Networks. IEEE Transactions on Wireless Communications 11(10), 3474–3483 (2012)

[3] Forouzan, B.A.: Data Communications and Networking, 3rd edn. Mc Graw Hill

[4] Islam, M.S., Sadid, W.H., Kashem, M.A., Rahman, M.A., Islam, M.N.: Wireless TCP in Unfair Situation: Performance Degradation and Improvement. In: Proc. of the International Multi Conference of Engineers and Computer Scientist (IMECS), vol. I, pp. 407–410 (March 2009)

[5] Bilel, B.R., Navid, N., Bouksiaa, M.S.M.: Hybrid CPU-GPU Distributed Framework for Large Scale Mobile Networks Simulation. In: Proc. 2012 IEEE/ACM 16th International Symposium on Distributed Simulation and Real Time Applications (DS-RT), pp. 44–53 (October 2012)

[6] Sanders, J., Kandrot, E.: CUDA by Example: An Introduction to General-Purpose GPU Programming, 1st edn. (July 29, 2010), http://www.amazon.com/CUDA-Example-Introduction-General-Purpose-Programming/dp/0131387685

[7] Kirk, D.B., Hwu, W.W.: Programming Massively Parallel Processors: A Hands-on Approach, 2nd edn. (December 28, 2012), http://www.amazon.com/Programming-Massively-Parallel-Processors-Edition/dp/0124159923/ref=dp_ob_title_bk

High-Performance Sound Engine of Guitar on Optimal Many-Core Processors

Myeongsu Kang[1], Cheol-Hong Kim[2], and Jong-Myon Kim[1,*]

[1] School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
{ilmareboy,jmkim07}@ulsan.ac.kr
[2] School of Electronics and Computer Engineering,
Chonnam National University, Kwangju, South Korea
cheolhong@gmail.com

Abstract. This paper presents design space exploration of optimal many-core processors for physics-based sound synthesis of an acoustic guitar by quantitatively evaluating the impact of the sample-per-processing element (*SPE*) ratio, which is the amount of sample data directly mapped to a processing element (PE). This paper evaluates system performance in terms of execution time, area and energy efficiencies for high-performance sound engine of the guitar as the *SPE* ratio is varied. Experimental results indicate that the *SPE* ratio in the range of 2,756 (or PEs=24) to 11,025 (or PEs=96) provides the most efficient operation for synthesizing guitar sounds with 6-note polyphony sampled at 44.1 *kHz*.

Keywords: Area efficiency, design space exploration, energy efficiency, many-core processors, physics-based sound synthesis, sample-per-processing element.

1 Introduction

Physical modeling synthesis has received increasing attention for creating high-quality sounds that imitate the sounds of natural instruments [1]. However, the computational complexity in physical modeling synthesis has limited its use in real-time applications due to the numerical integration of wave equations. To solve this problem, Motuk *et al.* investigated the use of field-programmable gate array to implement physical modeling synthesis and Luong *et al.* synthesized sounds of the gayageum, which is a Korean traditional plucked-string instrument, on a parallel processor by exploiting massive parallelism inherent in it [2, 3]. While it is evident that the overall performance improvement is achieved with increasing the number of processing elements (PEs), no general consensus has been reached that what granularity of processors and memories on the array system offers the most efficient performance for physical modeling synthesis. Consequently, this paper introduces a sample-per-processing element (*SPE*) which is the amount of sample data directly mapped to a PE, and explores the impact of *SPE* variations for physical modeling synthesis on system performance and efficiency characteristics.

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_89,

The rest of this paper is organized as follows. Section 2 presents simulation environment for evaluation the effect of different *SPE* configurations. Section 3 analyzes execution performance, area efficiency and energy efficiency for each *SPE* configuration. Section 4 concludes this paper.

2 Simulation Environments

2.1 Physical Modeling Synthesis: DWTVPS

Fig. 1(a) shows a plucked string instrument model and the waveguide is equivalent to a physical string. The plucked string instrument model is merged and rearranged due to linearity as illustrated in Fig. 1(b). Thus, a body "*impulse response*" is utilized as the string excitation to generate acoustic guitar sounds.

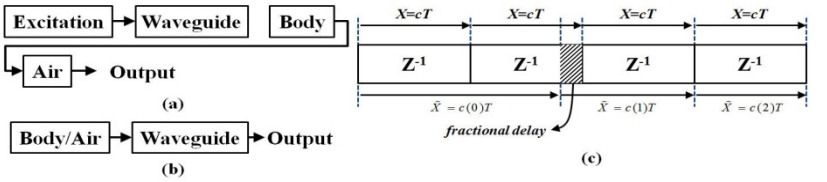


Fig. 1. Schematic diagram of the plucked string instrument and propagation distance corresponding to the propagation speed

To model physical strings, the first-order bidirectional linear digital waveguide is used in this study and its output for a discrete time constant n at position k is as follows [4]:

$$y(n,k) = y_r(n-k) + y_l(n+k), \tag{1}$$

where $y_r(n)$ and $y_l(n)$ are right/left-going traveling waves, respectively. A distance between discrete time indexes n and n+1 is denoted by X=cT and is represented as a unit delay as the propagation speed is uniformly distributed in space. If the propagation speed varies in time, the wave moves a distance $\tilde{X} = c(n)T$. Fig. 1(c) presents propagation distances with varying propagation speeds. For $c(n)>c$, the traveling waves move further than a unit delay while the traveling waves move shorter than a unit delay for $c(n)<c$. As depicted in Fig. 1(c), a fractional distance can occur and there is no way to directly express the fractional distance in digital domain. To deal with this problem, this paper employs a fractional delay filter as follows:

$$h(n) = \prod_{\substack{k=0 \\ k\neq n}}^{N} \frac{D-k}{n-k}, \quad n = 0,1,2,...,N, \tag{2}$$

where D is the fractional delay. To explain the digital waveguide model with time varying propagation speed (DWTVPS), the right-going traveling wave is only considered in this study, which can be expressed by $y(n,k)=y_r(n,k)$, since the left-going

traveling wave is explained in a similar manner to the right-going traveling wave. Given an initial distribution as $y(0,k)$, $k=-\infty,\ldots,\ \infty$, the outputs at an observation point k at n=1 and 2 are expressed by $y(1,k)=y_r(0,\ k\text{-}c(0)T)$ and $y(2,k)=y_r(0,\ k\text{-}c(0)T\text{-}c(1)T)$, respectively. Once the time varying velocity is defined as $c(n) = i(n) + c_f(n)$, where $i(n)$ is an integer value which is greater and equal to zero and $c_f(n)$ is a real value in [0, c], the output is generalized as

$$y(n,k) = y_r\left(0, k - \sum_{l=0}^{n-1} c(l)T\right) = y_r\left(0, k + M(n) + d(n)\right), \quad n > 0, \tag{3}$$

where $M(n) = -\sum_{l=0}^{n-1} i(l)$ and $d(n) = -T\sum_{l=0}^{n-1} c_f(l)$. $M(n)$ is a parameter related to the position of the traveling wave in the digital waveguide and $d(n)$ represents the fractional distance as shown in Fig. 1(c). The final output of DWTVPS is given as

$$y(n,k) = y_r\left(0, k + M(n) + d(n)\right) + y_l\left(0, k - M(n) - d(n)\right), \quad n > 0, \tag{4}$$

Despite the fact that it is possible to generate acoustic guitar sounds using DWTVPS, it is not possible to describe the frequency-dependent damping of a physical string. To represent frequency-dependent damping by dispersive wave propagation, a loop filter is needed as follows:

$$H_{loop}(z) = g\frac{1 + a_1}{1 + a_1 z^{-1}} \tag{5}$$

where g is the gain and a_1 is the filter coefficient which determines the cut-off frequency of the filter. To understand how each PE synthesizes desired guitar sounds, a pictorial representation of the full synthesis mechanism with a baseline many-core processor is illustrated in Fig. 2.

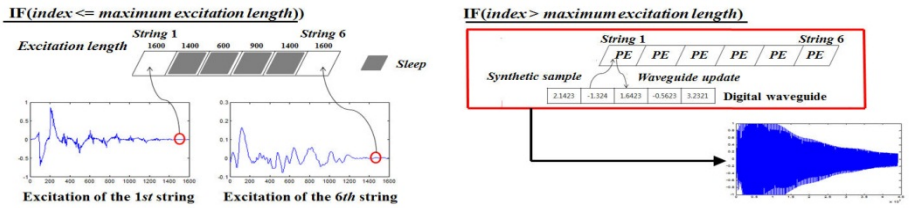


Fig. 2. Full mechanism for sound synthesis of the guitar

The many-core processor creates one-second-long 6-note polyphonic guitar sounds with a sampling rate of 44.1 *kHz* (i.e., 44,100 6-note polyphonic sounds samples are generated per second). In Fig. 2, the term "*index*" is an index value to store synthesized sound sample into the local memory of each PE. If the index value is less than equal to the length of the excitation signal, guitar sounds are generated with the corresponding excitations. In contrast, if the index value is greater than the length of the excitation signal, guitar sounds are synthesized by utilizing a previous sound sample in the waveguide, and then the waveguide is update with a synthetic sample. This process repeats until 44,100 sound samples are generated.

2.2 Many-Core Array Architecture

Fig. 3 illustrates the microarchitecture of the many-core processor array along with its interconnection network. It consists of a two-dimensional PE array, local memory and an array control unit (ACU). PEs are interconnected in a mesh structure; ACU controls the PE array; and an autonomous data exchange interface transfers data between an addressed word of memory in each PE and an array in the I/O unit's memory map. Each PE has a reduced instruction set computer datapath with the following minimum characteristics:

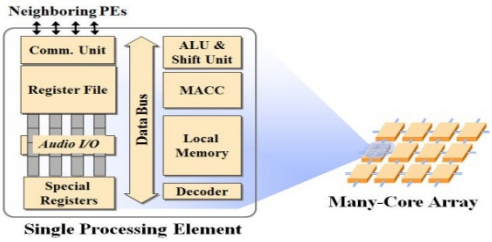


Fig. 3. A block diagram of a many-core array and a single processing element

- Small amount of 32-bit word local storage,
- Three-ported general-purpose registers (16 32-bit words),
- ALU computes basic arithmetic and logic operations,
- Barrel shifter performs multi-bit logic/arithmetic shift operations,
- Sleep unit activates or deactivates a PE based on local information,
- MACC multiplies 32-bit values and accumulates into a 64-bit accumulator, and
- Nearest neighbor communications through a north-east-west-south (NEWS) network and serial I/O unit

2.3 Methodology Infrastructure

A methodology infrastructure is divided into three levels: application, architecture, and technology. At the application level, single-instruction multiple-data (SIMD) parallel architecture simulator is used to profile execution statistics such as cycle count, dynamic instruction frequency and PE utilization for different *SPE* configurations by retargeting and optimizing physical modeling synthesis for each configuration based on the architecture and its execution properties. At the architecture level, the architectural modeling of functional units [5] for many-core arrays is utilized to calculate design parameters of each *SPE* configuration. The design parameters are then passed to the technology level. At the technology level, the generic system simulator (GENESYS) [6] is used to compute technology parameters such as latency, area, power and clock frequency for each *SPE* configuration. Finally, a design space explorer collects and combines all information such as cycle times, instruction latencies, instruction counts, area and powers of function units which obtained from the application, architecture and technology levels in order to determine execution times, area and energy efficiencies for each *SPE* configuration.

2.4 Design Space Exploration: *SPE* Variation

A key design issue for physical modeling synthesis is to determine the impact of direct access to sound sample data on processor element granularity. *SPE* variation is a selected design variable to determine the effect of grain sizes on the reference many-core processor architecture in this study. A discrete set of *SPE* ratio is defined as presented in Table 1 and its local memory size is calculated as follows:

$$MEM_{PE} = \frac{W_{MAX}}{PE_{STR}} + 4 \quad [words], \tag{2}$$

where PE_{STR} is the number of processing elements per each string, W_{MAX} is the maximum waveguide length to synthesize the guitar strings, which is set to 535. Likewise, an additional 4-word memory space is required to store filter coefficients: a loop filter coefficient, a loop gain, and fractional delay filter coefficients. As a result, an overall system memory is calculated by $MEM_{SYS} = MEM_{PE} \times N_{PE}$ [*words*], where N_{PE} is the number of PEs. In this study, seven *SPE* configurations are used as defined by $SPE=N_{sample}/PE_{STR}$, where N_{sample} is total number of samples to be generated. The number of PEs required for any given *SPE* configuration to cover the same sound sample size is given by $N_{PE}=6\times 2^{i}$, where i=0,…,6. For all configurations, a 44,100 sound data size is produced with a sampling rate of 44.1 *kHz* and 16-bit quantization. Moreover, all the configurations are implemented in 130 *nm* CMOS technology and 100 *MHz* clock frequency for performance analysis.

Table 1. Modeled many-core processor system parameters

Parameters	Values						
N_{PE}	6	12	24	48	96	192	384
SPE ratio	44,100	22,050	11,025	5,513	2,756	1,378	689
MEM_{PE} [*words*]	540	273	139	72	39	22	14
MEM_{SYS} [*KB*]	12.66	12.80	13.04	13.50	14.63	16.50	21

3 Experimental Results

3.1 Evaluation Metrics

Table 2 summarizes evaluation metrics for sound synthesis of the. Execution time is running time for sound synthesis, area efficiency is the amount of task throughput per unit of area, and energy efficiency is the task throughput achieved per *Joules*.

Table 2. Summary of evaluation metrics

Execution time	Area efficiency	Energy efficiency
$t_{exec} = \frac{C}{f_{clk}}$	$\eta_A = \frac{1}{t_{exec} \times Area}\left[\frac{1}{s \cdot mm^2}\right]$	$\eta_E = \frac{1}{t_{exec} \times Energy}\left[\frac{1}{s \cdot Joules}\right]$

where C is the cycle count, f_{clk} is the clock frequency, t_{exec} is the cycle time, *Area* is the PE array (mm^2), and *Energy* is the system energy required to complete the sound synthesis algorithm in 130 *nm* technology.

3.2 Synthetic Sounds and Execution Time

Fig. 4 shows spectra comparisons between synthesized sound using the many-core processor and original sounds, and they are very similar. Due to the limited number of pages allowed, more details about synthetic sounds are available at http://eucs.ulsan.ac.kr/FC2013/SoundEngine.

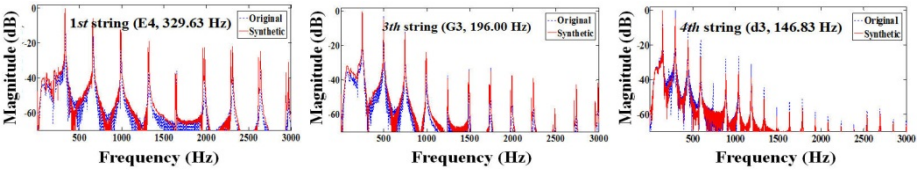


Fig. 4. Spectra of the original guitar sound (dotted line) and synthesized guitar sound using the reference many-core processor (solid line)

Table 3 presents execution times for variable *SPE* configurations. As expected, the execution time decreases as the number of PEs increases (or *SPE* ratio decreases) due to an increase in parallelism.

Table 3. Execution times for differnt *SPE* configurations

PEs	6	12	24	48	96	192	384
SPE ratios	44,100	22,050	11,025	5,513	2,756	1,378	689
Execution time [*ms*]	19.232	12.748	7.393	4.371	2.781	2.453	2.125

To obtain CD-quality sound sampled at 44.1 *kHz*, a sound sample must be synthesized within 1/44,100 seconds (about 0.2 *ms*). As shown in Table 3, for example, it takes 19.232 *ms* and 2.125 *ms* to synthesize 44,100 6-note polyphonic sound samples as the number of PEs is 6 and 384 (or *SPE* ratio is 44,100 and 689), respectively. This means that a sound sample for each guitar string can be synthesized within 0.436 μs and 0.048 μs, respectively. Thus, all of SPE configurations are fast enough to guarantee CD-quality sound.

3.3 Area and Energy Efficiencies

Area and energy efficiencies for different SPE configurations are illustrated in Fig. 5. The efficiencies are normalized to the task average for the purpose of comparisons. Consequently, the horizontal axis is not significant while the shape of the curve is significant.

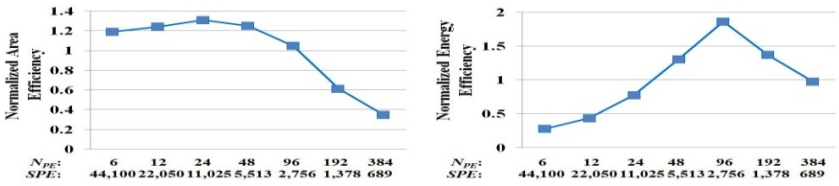


Fig. 5. Area energy (left) and energy efficiency (right) for variable *SPE* configurations

The maximum area efficiency is achieved at PEs=24 (*SPE* ratio=11,025) and this result correlates well with the combination of execution time (see Table 3) and system area (see Fig. 6). As shown in Table 3, the execution time is slightly decreased at above PEs=96 due to inter-PE communication operations in order to obtain updated sound samples between neighboring PEs. On the other hand, system area is exponentially increased at above PEs=24 since the system area is dominated by MACC. From these results, we can expect that the maximum are efficiency is achieved at PEs=24, which is the same as the simulation result. The maximum energy efficiency is achieved at PEs=96 as depicted in Fig. 5. According to Table 2, this result also correlates with the combination of execution time (see Table 3) and energy consumption (see Fig. 6). As shown in Table 3 and Fig. 6, execution time is slightly decreased at above PEs=96 while energy consumption is linearly increased to carry out additional processes such as transferring sound sample data between neighboring PEs.

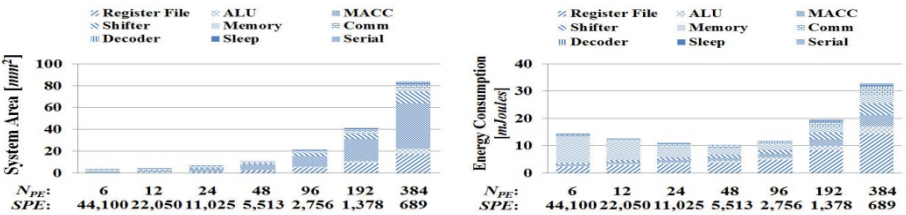


Fig. 6. System area (left) and energy consumption (right) for variable SPE configurations

3.4 Performance Comparison with Other Commercial Processors

Table 4 shows the performance comparisons between the selected optimal many-core processors (PEs=24 and 96) and other commercial processors (TI TMS320C6414, ARM 926EJ-S and ARM 10E). Although comparisons between the many-core processors and other commercial processors carry unavoidable errors, the goal of this study is to show the potential for improved performance of the optimal many-core processor for sound synthesis of the guitar, rather than a precise performance comparison. Experimental results show that the optimal many-core architecture with both PEs=24 and 96 outperforms other commercial processors in terms of area and energy efficiency.

Table 4. Performance comparison between the optimal many-core configuration and commercial processors

Parameters	ARM 926EJ-S	ARM 10E	TI TMS320C6414	Many-Core (PEs=24)	Many-Core (PEs=96)
Technology [*nm*]	130	130	130	130	130
Clock frequency [*MHz*]	250	400	720	100	100
Execution time [*ms*]	61.21	31.56	34.54	7.39	2.78
Energy consumption [*mJoule*]	7.34	6.31	32.8	10.85	11.79
Area [mm^2]	2.78	10.3	529	6.41	21.3
Energy efficiency [$1/s \cdot Joules$]	2226	5022	883	12471	30509
Area efficiency [$1/s \cdot mm^2$]	5.88	3.08	0.05	10.56	8.44

4 Conclusions

The design space of optimal many-core processors for the physics-based sound synthesis was explored by quantitatively evaluating the impact of the *SPE* ratio on system performance and efficiency. The analysis presented indicated that the *SPE* ratio in the range of 2,756 and 11,025 (or the number of PEs between 24 and 96) provides the most efficient many-core architecture that maximizes performance per cost and energy for synthesizing guitar sounds with 6-note polyphony sampled at 44.1 *kHz* with 16-bit quantization. In addition, the selected optimal many-core architecture achieves 211-fold (PEs=24) and 168-fold (PEs=96) performance improvements in area efficiency, 14-fold (PEs=24) and 34-fold (PEs=96) performance improvements in energy efficiency over the commercial TMS320C6414 processor, respectively.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566).

References

1. Serra, X.: State of the Art and Future Directions in Musical Sound Synthesis. In: IEEE 9th Workshop on Multimedia Signal Processing, Crete, pp. 9–12 (2007)
2. Motuk, E., Woods, R., Bilbao, S.: FPGA-Based Hardware for Physical Modeling Sound Synthesis by Finite Difference Schemes. In: International Conference on Field-Programmable Technology, Singapore, pp. 103–110 (2005)
3. Van Luong, H., Cho, S., Kim, J.M., Chong, U.: Real-Time Sound Synthesis of Plucked String Instruments Using a Data Parallel Architecture. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS, vol. 5754, pp. 211–221. Springer, Heidelberg (2009)

4. Smith, J.O.: Physical Modeling Using Digital Waveguides. Comput. Music J. 16(4), 74–91 (1992)
5. Chai, S.M., Taha, T., Wills, D.S., Meindl, J.D.: Heterogeneous Architecture Models for Interconnected-Motivated System Design. IEEE Trans. VLSI Syst. 8(6), 660–670 (2000)
6. Nugent, S., Wills, D.S., Meindl, J.D.: A Hierarchical Block-based Modeling Methodology for SoC in GENESYS. In: International ASIC/SOC Conference, New York, pp. 239–243 (2002)

Community Identification in Multiple Relationship Social Networks

Ting-An Hsieh[1], Kuan-Ching Li[2], Kuo-Chan Huang[1], Kuo-Hsun Hsu[1], Ching-Hsien Hsu[3], and Kuan-Chou Lai[1,*]

[1] Depart. of Computer Science, National Taichung University, Taiwan
[2] Depart. of Computer Science and Information Engineering, Providence University, Taiwan
[3] Depart. of Computer Science Information Engineering, Chung Hua University, Taiwan

Abstract. In a social network, individuals often simultaneously belong to multiple social communities; therefore, the detection of relationships among individuals is very important. However, most of community detection methods only apply a single relationship in dynamic social networks with multi-relationships among individuals. Therefore, this study proposes a CNET Hierarchical Division Algorithm (CHDA) to detect communities efficiently. Experimental results show that the proposed CHDA could detect communities with more precise recognition, regarding their characterization.

Keywords: community detection, dynamic social network, hierarchical division, recognition.

1 Introduction

Analyzing social networks is a quantitative analysis approach; the sociologist uses the graph theory and other mathematical methods to analyze the social behavior. Recently, the focus of developing community detection algorithms has shifted from the technique of cluster classification to that of the characterization by sharing resource properties, so the detection of relationships among individuals is a very important issue. In general, individuals often simultaneously belong to multiple social communities. If the community detection in dynamic social networks adopts the hierarchical division algorithm, some information may be missed. Most of current community detection approaches only consider one single relationship in the dynamic social network; however, in a real world, there are multiple relationships among individuals in the dynamic social network. The clique percolation approach is an analyzing strategy for the overlapping community networks [2]. This proposed approach finds the communities from k-cliques, which correspond to complete (fully connected) sub-graphs of k nodes. A community is presented as the maximal union of k-cliques that can be reached from each other through a series of adjacent k-cliques. A node with a large degree of the major significances is an individual with a high connection degree. In such a case, it

* Corresponding author.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_90,

may not express the join of a community, so the node identification with a large edge degree cannot find the correct community [3]. In the same situation, a node with a high connectivity is unable to identify communities.

This paper proposes a Cell Net (CNET) Hierarchical Division Algorithm (CHDA) to identify the community. The proposed approach could identify the multi-relationships in the dynamic social network without missing any information. The rest of this study is shown as follows. Section 2 presents some of the related works. The CHDA, its experimental environment and the experimental results are discussed in Section 3 and 4. Section 5 gives some conclusions.

2 Related Works

Previous related work [5] points out the six types of dynamic social networks, and predicts the user behavior by using the information. An efficient algorithm is proposed for detecting and tracking community dynamics based on the graph topology and community characteristics. The study [2] observes the degree distribution in networks by using the preferential attachment model. According to this model, the authors adopt a dynamical hypothesis for detecting the communities.

Most of the community detection approaches adopt the clique theory [2][4] or hierarchical division approaches [1] for identifying communities. The previous work [4] tries to find the maximal clique and to group the overlapped cliques to be a clustering kernel. The proposed strategy deals with the agglomerative process according to the proposed distance measurement, and assigns the nodes to their closest kernel. In another study [1], each node is a community initially, and the proposed approach applies the local similarity measurement and the diffusion of membership to different communities. Some previous studies [2][7] consider that community overlapping may hide information from networks. The study [2] tries to detect communities in large-scale social networks by using communities' overlapping nature. The authors in the work [7] propose a new model to represent an interconnected network of networks. The proposed model shows the topology of on-line social networks, and the interaction in different networks at the same time.

3 CNET Hierarchical Division Algorithm

A dynamic social network is formed by a number of communities, and each community consists of a number of cliques. In general, each clique is a cell network (CNET), and a CNET is one of the largest complete connection graphs. So, a clique is a community, but the community is not necessary to be a clique. Suppose the maximum range of a dynamic social network is R, and the smallest unit of distance in the coordinate system is ω. The number of divisions equals to [log(R/ ω)]. The position of each node in the dynamic network depends on the close degree between cliques.

The proposed algorithm consists of five steps. Fig. 1 shows an example of hierarchical cells. Firstly, the dynamic network is divided into a hierarchy of cells, in which each cell at any intermediate level is divided into a 2*2 grid of four equal-size sub-cells

at the next lower level. The partitioning procedure proceeds until the cell size at the bottom level is within the minimum distinguishable unit of the dynamic network.

Secondly, the cliques are listed as $CNET_1$, $CNET_2$,…, and $CNET_i$. There are four cases between any two CNETs. The first case shows the overlapping at the same level in different cells between two CNETs. The second case presents the overlapping at the same level in the same cell between two CNETs. The third case shows the separation at the same level in the same cell between two CNETs. The last case presents the separation at the same level in different cells between two CNETs.

The third step merges CNETs with satisfying following conditions: First, when the number of common nodes is N, where N= V*α, V is the number of nodes, α=2/V. V is the total number of the nodes initially. When we construct a community between two cliques with only one common node compared with two or more than two common nodes, their relationship is relatively weak. Second, the distance between CNETs, D, where $0 \leqq D \leqq [V*\alpha]+[n*\beta]$, and β=1/n. n is the dynamic network range initially. The distance between two cliques is 1 composed of a community compared with the distance between cliques more than one, and the relationship is strong.

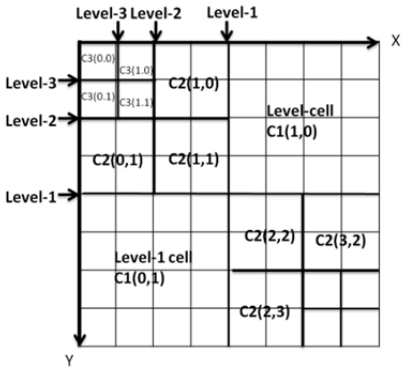


Fig. 1. Example of the hierarchical cells

As new nodes join, the network continues to expand its scope without changing α and β. Only the number of nodes (V) and the range (n) of the network are changed. The last step checks whether there is any overlap between CNETs on the same level. If there is overlapping and the third condition is satisfied, the proposed approach merges these communities; otherwise, each CNET could be a community. In the last step, the proposed approach identifies the node which cannot be included in any CNET, and determines its belonging the community by the location function.

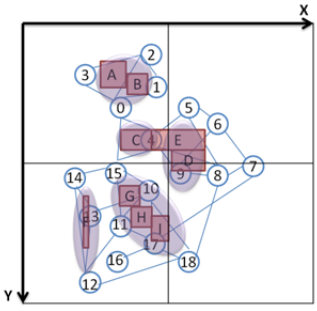
Algorithm: CNET Hierarchical Division Algorithm
Input: nodes in a multi-relationship social network
Output: communities in the multi-relationship social network
{list CNETs in the multi-dimension space
While (∀ Node ∈ SN)
If (there is a complete graph) identify to be CNET

determine the relationship between CNETs in the same level
Search_Level = [(the dynamic social network range)/(log2)]
For (determined level is 1; determine level < search level; determine level ++) {
calculate V by summing all of node counts, and the distance between CNETs, D
calculate α to be 2/V
let β be 1/(the dynamic social network range)
If (CNET's level in the determine Level & there is the same node of the N number & $0 \leqq D \leqq [V*\alpha]+[n*\beta]$)
CNETs is in the same community
the cell is divided into 2*2 cells }
identify the node that is not in communities
Let NODES be the set of nodes that are not in communities
For each node in NODES {
Calculate array DISTS to be the distance between node and CNETs
For each distance in DISTS {
find the minimum distance
let the node be merged into the community } }}

4 Experiment Results

The proposed algorithm is demonstrated by the following example. Assume that the grid size of the dynamic network is 8 * 8, the initial number of nodes is 18, and the smallest unit of a coordinate system is 1. Therefore, the number of division is 3 ($[(\log 2^3/2^0)]$). When there is no new joining nodes and two cliques tries to merge, these two cliques have to satisfy the following conditions: the common nodes N equals to V*2/18, and the distance D between CNETs satisfies $0 \leqq D \leqq [V*1/9]+[n*1/8]$.

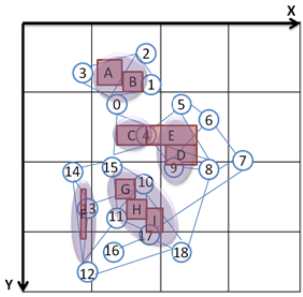
In the first division, the hierarchical cells in the 2*2 grid are shown in Fig.2. The proposed approach identifies $CNET_D$ and $CNET_E$ at the same level in different cells with overlapping, and finds the satisfactions of the conditions. The proposed approach merges them into a community.



CNET	Location	Node	
A	(0,0)	0,2,3	Community
B	(0,0)	0,1,2	
C	(0,0)	0,4,15	Community
D	(1,1)	4,8,9	Community
E	(1,0)	4,5,8	
F	(0,1)	12,1,3,14	Community
G	(0,1)	10,11,15	Community
H	(0,1)	10,11,17	
I	(0,1)	10,17,18	

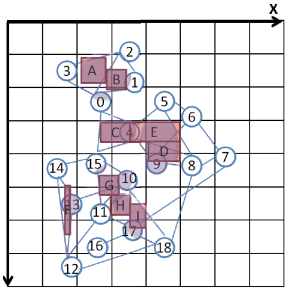
Fig. 2. First division

In the second division, the hierarchical cell of a 4*4 grid and the community relations are shown in Fig.3. In this phase, no community is separated. The hierarchical cell of an 8*8 grid and the community relations are shown in Fig.4. In this phase, $CNET_G$ consists of the community H and I. In each division, the proposed approach generates close degree tables, as shown in Table 1.



CNET	Location	Node	
A	(1,0)	0,2,3	Community
B	(1,0)	0,1,2	
C	(1,1)	0,4,15	Community
D	(2,1)	4,8,9	Community
E	(2,0)	4,5,8	
F	(0,2)	12,1,3,14	Community
G	(1,2)	10,11,15	Community
H	(1,2)	10,11,17	
I	(1,3)	10,17,18	

Fig. 3. Second division



CNET	Location	Node	
A	(2,1)	0,2,3	Community
B	(3,1)	0,1,2	
C	(2,3)	0,4,15	Community
D	(4,4)	4,8,9	Community
E	(4,3)	4,5,8	
F	(1,6)	12,1,3,14	Community
G	(2,4)	10,11,15	Community
H	(3,5)	10,11,17	Community
I	(3,5)	10,17,18	

Fig. 4. Third division

Table 1. Generated close degree tables

(a) first division

Division	Community	Close degree
1	AB	1
1	C	1
1	DE	1
1	F	1
	GHI	1

(b) second division

Division	Community	Close degree
2	AB	2
2	C	1
2	DE	2
2	F	1
2	GHI	2

(c) third division

Division	Community	Close degree
3	AB	3
3	C	1
3	DE	3
3	F	1
3	G	3
3	HI	3

In the last step, in order to identify node v6、v7、v16 which cannot be included in any CNET's, the proposed approach determines the belonging communities by using the location function. Node v6 only has the connection with community D and E, so that v6 is included into these communities. The distance between node v7 (6, 3) and community D and E (4, 4) is 5, and the one between community H and I (3, 5) is 13, so node v7 is included into community D and E. Node v16 has the connection only with community H and I, so it is included into this community.

With the close degree tables, the community A and B, community D and E, community H, I could be identified. After the third division, they still belong to the same community.

5 Conclusions and Future Work

In this study, we propose a community detection strategy for the multi-relationship social network. Individuals often span across many different social networks, such as working environments, league, or club activities and so on. When the dynamic social network is handled by the hierarchical division algorithm, it may miss some interaction information in different social groups at the same time. However, our approach could avoid such a situation. In the future, our work will focus on the CNET hierarchical division algorithm for detecting the community and finding the impact of personality in the dynamic social network.

Acknowledgment. This study was sponsored by the National Science Council, Taiwan, the Republic of China, under contract numbers: NSC 101-2221-E-142-001- and Delta Electronics Inc. under contract number: 101F2289A8.

References

1. Cazabet, R., Amblard, F., Hanachi, C.: Detection of Overlapping Communities in Dynamical Social Networks. In: 2010 IEEE Second International Conference on 2010 Social Computing (SocialCom), pp. 309–314. IEEE (2010)
2. Palla, G., et al.: Uncovering the overlapping community structure of complex networks in nature and society (2005)
3. Du, N., et al.: Community Detection in Large-Scale Social Networks.pdf. In: 9th WebKDD and 1st SNA-KDD 2007 Workshop. ACM, New York (2007)
4. Chen, Z., et al.: Detecting and Tracking Community Dynamics in Evolutionary Networks (2010)
5. Rees, B.S., Gallagher, K.B.: Overlapping Community Detection by Collective Friendship Group Inference, pp. 375–379 (2010)
6. Magnani, M., Rossi, L.: The ML-Model for Multi-layer Social Networks. In: Advances in Social Networks Analysis and Mining 2011, pp. 5–12. IEEE, Kaohsiung (2011)
7. Goldberg, M., et al.: Finding Overlapping Communities in Social Networks, pp. 104–113 (2010)

An Improved ACO by Neighborhood Strategy for Color Image Segmentation

Shih-Pang Tseng[1,3], Ming-Chao Chiang[1], and Chu-Sing Yang[2]

[1] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, R.O.C.
[2] Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.
[3] Department of Computer Science and Information Engineering, Tajen University, Pingtung, Taiwan, R.O.C.
tsp@mail.tajen.edu.tw, mcchiang@cse.nsysu.edu.tw, csyang@ee.ncku.edu.tw

Abstract. This paper presents an efficient method for speeding up ant colony optimization (ACO) in solving the color image segmentation problem. The proposed method is inspired by the heuristics of image segmentation to reduce the computation time. To evaluate the performance of the proposed method, we applied the method on well-known test images. Our experimental results shows that the proposed method can significantly reduce the computation time about 19% to 45%.

Keywords: Color image segmentation, clustering, ant colony optimization.

1 Introduction

Image segmentation is the process of partitioning the pixels of one image into multiple clusters which contains the pixels with the similar visual characteristics [1]. In image processing and computer vision, image segmentation is one important problems since it is the pre-processing work of many other advanced tasks. In the past, many kinds of methods are proposed for image segmentation problem. Clustering-based methods for image segmentation is an important research direction [2][3].

In general, there are two types of image segmentation approaches, one is region-based and the other is contour-based. [4] Region-based usually approaches find the regions of pixels with similar characteristics such as color. In contrast, contour-based approaches try to detect the possible edges and to divide the image into different regions by these edges.

The basic idea of the region-based approaches is that the pixels in the identical image region should have some similar characteristics. Therefore the region-based approaches usually use data clustering techniques, such as nearest-neighbor K-means, Fuzzy C-means, and so on.

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_91,

The remainder of the paper is organized as follows. The related works are in Section 2. Section 3 introduced the ACO and our improvement. Performance evaluation is presented in Section 4. Conclusion is given in Section 5.

2 Related Work

In the last 20 years, the evolutionary computing has successfully been applied on various discrete and continuous problems. Bhanu et. al[5] used genetic algorithm as a image segmenting controller to improve the quality of segmentation. Another approach is directly using evolutionary clustering algorithm. For example, Belahbib et. al [6] combined genetic algorithm and Fuzzy C-means to the image segmentation. Particle swarm optimization (PSO) is another popular evolutionary clustering methods, such as Chander et. al[7] proposed a adaptive PSO variant for image segmentation. Ant colony optimization (ACO) is a probabilistic technique for discrete problems including data clustering and image segmentation. Liang et. al[8] proposed a hybrid approach based on an ACO with the Otsu method. But Liangs' approach only uses the pheromone information and not uses the heuristic information. This means that more iterations is needed for searching process. Tai et. al[9] proposed an ACO for image segmentation with fuzzy entropy which guides the local and global search process. There are some variants of ACO, such as MAX-MIN ant system (MMAS) [10], proposed to improve the effectiveness and efficiency. MMAS simply modified the pheromone update rule to avoid premature convergence and improve the search efficiency. The maximum and minimum pheromone value are set to avoid premature convergence. Only the best ant can trail pheromone to improve the search efficiency. That we use the MMAS as the typical ACO in this paper.

3 Proposed Method

The *SolutionConstruction()* and *PheromoneUpdate()* are two major process of ACO. In the *SolutionConstruction()* process, each ant constructs its solution by the pheromone and heuristic information. Let $x_i \in C_j$ denote the pixel x_i is the member of the cluster C_j. In the *SolutionConstruction()* process, the ant assigns the x_i to C_j with the following probability:

$$p_{ij} = \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{j=1}^{M}\tau_{ij}^{\alpha}\eta_{ij}^{\beta}} \tag{1}$$

$$\eta_{ij} = \frac{1}{\|x_i - m_j\| + 1} \tag{2}$$

where M is the number of clusters, τ_{ij} is the pheromone value between the pixel x_i and cluster C_j. m_j is the mean of cluster C_j. η_{ij} is the heuristic value, and η_{ij} is the inverse of the Euclidean distance originally. To avoid that the distance may be zero, we do a slight modification.

The pheromone values are represented the previous experience of ACO search process. The *PheromoneUpdate()* procecss update the pheromone values as following:

$$\begin{cases} \tau_{ij} = (1-\rho)\tau_{ij} + \Delta\tau & \text{if } x_i \in C_j \text{ in the best ant} \\ \tau_{ij} = (1-\rho)\tau_{ij} & \text{Others} \end{cases} \tag{3}$$

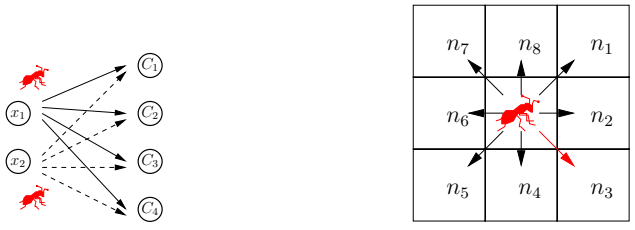
where $\rho \in (0, 1]$ denotes the evaporation rate.

3.1 Neighborhood Strategy

As the above description of ACO, each ant assign each pixel by the pheromone value τ and heuristic information η. The pheromone values are stored in the pheromone table. In the *SolutionConstruction()* process, to get the pheromone value is just to look up the pheromone table; and this takes a little computation time. But the η value is derived from the Euclidean distance between the pixel and the mean of the cluster, this would take a large amount of computation time. Furthermore, the means of the clusters are varied each iteration, the Euclidean distances are needed to evaluate each time. As shows in Fig. 1(a), each ant must do the evaluations of Euclidean distance $N \times M$ times in one iterations, where the N is the number of pixels and the M is the number of clusters. Let the θ denoted the number of ants and the l denoted the number of iterations. In the whole ACO process, there are $l \times \theta \times M \times N$ times of Euclidean distance evaluations.

In order to reduce the computation time of Euclidean distance, we proposes the neighborhood strategy applied on ACO. The ACO with neighborhood strategy is named Neighborhood ACO (NACO). It is based on the assumption that each pixel has at least one neighbor pixel which is belong to the same cluster in color image segmentation problem. And the neighbor pixel with smaller Euclidean distance has the larger probability in the same cluster. Thus, we used the neighbors to replace the means of clusters in the *SolutionConstruction()* process. The ant does not directly assign the pixel to the cluster. The ant decides which neighbor is in the same cluster with the pixel and indirectly assigns the pixel to the cluster. As shown in Fig. 1(b), a pixel has only eight neighbors in the image. And the Euclidean distance between neighbors are fixed, we can pre-compute the Euclidean distances and store in a table for the *SolutionConstruction()* process.

NACO has only $8 \times N$ times of Euclidean distance evaluations.



(a) (b)

Fig. 1. Example illustrating how each pixel is assigned to the cluster (a) directly (b) via its neighbor

4 Experimental Result

The experimental result was conducted on a HP DL165 G7 machine with 2.6 GHz AMD Opteron CPU and 12GB of memory using Ubuntu 12.04. Moreover, all the programs are written in C++ and compiled using g++ (GNU C++ compiler). As shown in Table 1, five images—denoted *Birds, Lake, Moon, Red Church*, and *Swimmer* —are used to measure the performance of the ACO and NACO. All five images are from the Berkeley Segmentation Dataset and Benchmark (`http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`). All images are of size 481 × 321 and in 24-bit *rgb* color space. The number of clusters based on the research by Tan et al. [3]

Table 1. Test images from

Image	Size	Color	#clusters
Birds	481 × 321	*rgb*	3
Lake	481 × 321	*rgb*	12
Moon	481 × 321	*rgb*	3
Red Church	481 × 321	*rgb*	10
Swimmer	481 × 321	*rgb*	12

To evaluate the quality of color image segmentation, the mean square error *(MSE)* is used in this paper. A better clustering method should generate results with smaller distortions than the worse one. The *MSE* evaluation function could be described as follows:

$$MSE = \frac{1}{N}\sum_{j=1}^{M}\sum_{i\in C_j}\|x_i - m_j\|^2 \tag{4}$$

To simplify the discussion of the experimental results, we will use the following conventions. Let *MSE* and *Time* denote, respectively, the quality of segmentation and the computation time. Let

$$\Delta MSE = \frac{MSE_{NACO} - MSE_{ACO}}{MSE_{ACO}} \times 100\% \tag{5}$$

$$\Delta Time = \frac{Time_{NACO} - Time_{ACO}}{Time_{ACO}} \times 100\% \tag{6}$$

where ΔMSE and $\Delta Time$ denotes the enhancement of NACO with respect to ACO in percentage. Because the segmentation method with smaller *MSE* or *Time* is better, a more negative value of ΔMSE or $\Delta Time$ implies a greater enhancement.

As far as this paper is concerned, for all the ACO and NACO we evaluated, the number of ants θ is set equal to 20, and the number of iterations l is set equal to 1,000. The other parameter settings, we referred to [11]. The α and β are set

to 1 and 2. The pheromone evaporation rate ρ is 0.9. The initial pheromone value τ_0 is set to 0.5. The pheromone increment $\Delta\tau$ is 0.1. The maximum and minimum pheromone value, τ_{MAX} and τ_{MIN} is set to 0.99 and 0.01. The initial solution is generated by randomizing. The K-means process is used as a local search operator in both ACO and NACO. Each experiment repeated 30 runs and we showed the average in Table 2 and Table 3.

Table 2. Comparison of clustering quality between the proposed NACO and ACO techniques based on the MSE

Image	ACO	NACO				
	MSE	*Time*	*MSE*	*Time*	ΔMSE	$\Delta Time$
Birds	183.9	2032.1	183.8	1627.2	-0.02%	-19.9%
Lake	407.5	6702.5	404.3	3680.1	-0.77%	-45.1%
Moon	304.1	2065.0	304.1	1662.6	0.02%	-19.5%
Red Church	185.4	5607.4	189.1	3178.1	1.97%	-43.3%
Swimmer	316.8	6697.6	317.6	3661.8	0.23%	-45.3%

The comparison of ACO and NACO is shown in Table 2. The five test images can be divided into two groups by the number of clusters. The first group is *Birds* and *Moon* because of the small number of clusters. The neighborhood strategy applied on ACO can reduce about 19% execution time in the color image segmentation with small number of clusters. The other group is *Lake*, *Red Church* and *Swimmer*. The NACO can reduce about 43% to 45% execution time. It shows that the neighborhood strategy can significantly improve the efficiency of ACO in color image segmentation. In addition, the difference of ACO and NACO is very small.

Table 3. Results showing percentage of computation time that can be reduced by NACO with different number of clusters

Image	#clusters	ACO		NACO			
		MSE	*Time*	*MSE*	*Time*	ΔMSE	$\Delta Time$
Lake	4	1276.9	2561.0	1276.6	1866.7	-0.02	-27.1
Lake	6	840.2	3589.9	836.4	2329.0	-0.45	-35.1
Lake	8	624.6	4589.2	623.6	2742.9	-0.16	-40.2
Lake	10	501.9	5726.4	499.0	3189.2	-0.59	-44.3
Lake	12	407.5	6702.5	404.3	3680.1	-0.77	-45.1
Lake	14	338.8	7606.1	337.1	4376.8	-0.50	-42.5
Lake	16	298.8	8658.1	299.4	4569.2	0.20	-47.2

In order to illustrate the relationship between reduction of execution time and the number of clusters, we applied the different number of clusters on the image *Lake*. Table 3 shows the percentage of computation time that can be reduced by NACO with different number of clusters in *Lake* image. In general, the reduction ratio increases by the number of clusters.

5 Conclusion

Image segmentation is an important problem in image processing and computer vision. In this paper, we proposed an improved ACO, named NACO, by neighborhood strategy. The strategy is1 based on the assumption that each pixel has at least one neighbor pixel which is belong to the same cluster in color image segmentation problem. In our analysis, the NACO can reduce a large amount of computation time. And our experimental result shows that the NACO is more efficient than ACO. One weakness of NACO is that the number of clusters should be pre-defined. In the future, we will extend the NACO to unsupervised image segmentation. And we will use the better local search operator, such as Fuzzy C-means, to improve the quality of segmentation in NACO.

References

1. Haralick, R.M., Shapiro, L.G.: Image segmentation techniques. Computer Vision, Graphics, and Image Processing 29(1), 100–132 (1985)
2. Yu, Z., Au, O.C., Zou, R., Yu, W., Tian, J.: An adaptive unsupervised approach toward pixel clustering and color image segmentation. Pattern Recognition 43(5), 1889–1906 (2010)
3. Tan, K.S., Isa, N.A.M., Lim, W.H.: Color image segmentation using adaptive unsupervised clustering approach. Applied Soft Computing 13(4), 2017–2036 (2013)
4. Malik, J., Belongie, S., Leung, T., Shi, J.: Contour and texture analysis for image segmentation. Int. J. Comput. Vision 43(1), 7–27 (2001)
5. Bhanu, B., Lee, S., Ming, J.: Adaptive image segmentation using a genetic algorithm. IEEE Transactions on Systems, Man and Cybernetics 25(12), 1543–1567 (1995)
6. Bellala Belahbib, F.Z., Souami, F.: Color image segmentation by a genetic algorithm based clustering and connected component labeling. In: 2012 24th International Conference on Microelectronics (ICM), pp. 1–4 (2012)
7. Chander, A., Chatterjee, A., Siarry, P.: A new social and momentum component adaptive pso algorithm for image segmentation. Expert Systems with Applications 38(5), 4998–5004 (2011)
8. Liang, Y.-C., Chen, A.H.-L., Chyu, C.-C.: Application of a hybrid ant colony optimization for the multilevel thresholding in image processing. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006. LNCS, vol. 4233, pp. 1183–1192. Springer, Heidelberg (2006)
9. Tao, W., Jin, H., Liu, L.: Object segmentation using ant colony optimization algorithm and fuzzy entropy. Pattern Recognition Letters 28(7), 788–796 (2007)
10. Stuützle, T., Hoos, H.H.: Maxmin ant system. Future Generation Computer Systems 16(8), 889–914 (2000)
11. Dorigo, M., Stuützle, T.: Ant Colony Optimization. The MIT Press (2004)

A Novel Spiral Optimization for Clustering

Chun-Wei Tsai[1], Bo-Chi Huang[2], and Ming-Chao Chiang[2]

[1] Department of Applied Informatics and Multimedia,
Chia Nan University of Pharmacy & Science, Tainan, Taiwan
cwtsai0807@gmail.com
[2] Department of Computer Science and Engineering,
National Sun Yat-sen University, Kaohsiung, Taiwan
bchuang0405@gmail.com, mcchiang@cse.nsysu.edu.tw

Abstract. Because most traditional search methods are unable to satisfy the current needs of data mining, finding a high performance search method for data mining has gradually become a critical issue. The spiral optimization (SO) is a promising search algorithm designed to emulate the natural phenomena, such as swirl and low pressure, to find the solutions of optimization problems within an acceptable computation time. In this paper, a novel SO is presented to solve the clustering problem. Unlike the original SO, which rotates the points around the elitist center iteratively, the proposed algorithm, called distributed spiral optimization (dSO), splits the population into several subpopulations so as to increase the diversity of search to further improve the clustering result. The *k*-means and oscillation methods are also used to enhance the efficacy of dSO. To evaluate the performance of the proposed algorithm, we apply it to the clustering problem and compare the results it found with those of the spiral optimization and genetic k-means algorithm. The results show that the proposed algorithm is quite promising.

Keywords: Metaheuristic, spiral optimization, and clustering.

1 Introduction

Clustering is a well-known optimization problem, partially because it is a hard problem in terms of the time complexity (NP-hard [1]) and partially because its relevant technologies can be applied to many real world problems, such as web document classification [2], face recognition [3], and the analysis of customer behavior [4]. Although the information technologies have advanced in recent years, finding a powerful clustering technique is still a challenge, especially when taking into account the big data problem we are facing nowadays.

In addition to the early nature-inspired algorithms, evolutionary algorithms, and swarm intelligence, a number of new metaheuristics have also been presented more recently, some of which are developed to enhance the search performance during the convergence process compared with early metaheuristics, such as harmony search [5], bees optimization [6], firefly algorithm [7], and spiral optimization [8]. Among them, the spiral optimization (SO) is a promising one because it is simple and easy to

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*,
Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_92,

implement. In general, the spiral optimization presented by Tamura and Yasuda was inspired by the analogy of spiral phenomena in nature [8]. The main idea of the spiral optimization is that all the points (individuals) rotate about the center (i.e., the elitist or optimal solution) to search the solution space, and the population converges to the best solution in the end. In the spiral model, all the points attracted by the center move to their next positions by using the rotation operator, the only operator of SO. Although SO outperforms the traditional single-solution-based algorithms for continuous optimization problems, the center of SO is still easy to fall into local minimum. For this reason, we present in this study a novel distributed spiral optimization to enhance the performance of SO for the clustering problem[1]—by splitting the population into a number of subpopulations and introducing two operators to the SO: the one-iteration k-means and oscillation operators.

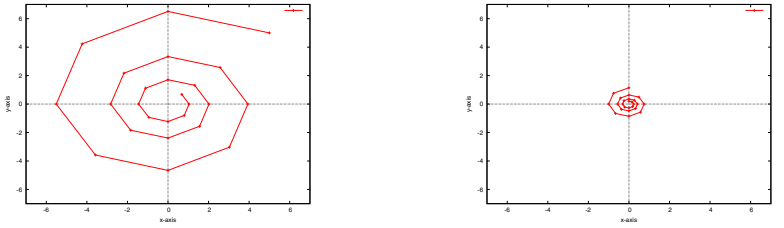
The rest of the paper is organized as follows. In Section 2, a brief introduction to the spiral optimization is given. After that, the proposed algorithm is detailed in Section 3. Section 4 compares the simulation results of the proposed algorithm with those of other clustering algorithms. The conclusion is drawn in Section 5.

2 Spiral Optimization

In this section, a brief review of the spiral optimization (SO) for continuous optimization problems [8] is given to show how it works. As noted previously, the main concept of the SO is to rotate the points about a center to find the potential solutions during the convergence process. As shown in Fig. 1, on one hand, in the early stage of the convergence process, SO will search for a large space to find regions that have a higher probability to find a better solution. In other words, SO will try to maintain the diversification in the early stage of the convergence process. On the other hand, in the later stage of the convergence process, SO will focus on finding better solutions in one of the regions found in the early stage [8]. In other words, SO will try to maintain the intensification in the later stage of the convergence process.

The spiral optimization is as outlined in Fig. 2. As shown in line 2, the rotation angle θ and the convergence rate r are set to, respectively, a value in the rotation angle$\theta$and the convergence rate r are set to, respectively, a value in the open interval $(0, 2\pi)$ and a value in the open interval $(0, 1)$. That is, $\theta \in (0, 2\pi)$ and $r \in (0, 1)$. Then, as shown in line 3, the population X is initialized as a set of centroids, i.e., $X = \{x_1, x_2, \ldots, x_N\}$, where N denotes the population size. The best point (individual) is then selected as the center (elitist) by a predefined measurement, such as the sum of squares error (SSE) for the clustering problem. Finally, in the main loop (also called the convergence process), the points will be rotated about the center by the angleθ, and the points is closer to the center by the convergence rate r. After all points finish performing the rotation operator, the population select the best fitness value in this generation as the new center. In the spiral model, the points is attracted by the center,

[1] The solutions of the clustering problem can be encoded either by the centroids (a continuous representation) or by the clusters to which patterns are assigned (a discrete representation).



(a) The search scope of the first 25 iterations (b) The search scope of the last 25 iterations

Fig. 1. The search strategy of the spiral optimization

and move to the next position regularly, as shown in line 7. The termination criterion then is check to see if the termination criterion met, the spiral optimization will stop and return the best-so-far point as the final result; otherwise, the spiral optimization will continue to perform the spiral model.

```
1  procedure SpiralOptimization() {
2      Set the parameters 0 < θ < 2π and 0 < r < 1.
3      Initialize the population of points x_i and select the best point x* as the center.
4      Let k = 0;
5      while (k < k_max)
6          for each point x_i
7              x_i(k + 1) = Rotation(x_i(k), x*, θ, r);
8          end for
9          x* = UpdateCenter();
10         k = k + 1;
11     end while
12 }
```

Fig. 2. Outline of the spiral optimization

The rotation operator, which plays an important role in deciding how to *move* the search point of the spiral optimization to the next position during the convergence process, will be discussed in detail below.

Consider the two-dimensional space. Fig. 3 gives an example to show how a point x in the two-dimensional space is rotated about the origin (0, 0) by an angle θ to x'; that is,

$$x' = R^{(2)}(\theta)\, x, \tag{1}$$

where the rotation matrix is defined as

$$R^{(2)}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2}$$

The rotation matrix $R^{(2)}(\theta)$ can be used to construct the spiral model in the two-dimensional space by adding the convergence rate r to Eq. (1), as follows:

$$x(k+1) = r\, R_{(2)}(\theta)\, x(k) = S_2(r,\theta)\, x(k), \tag{3}$$

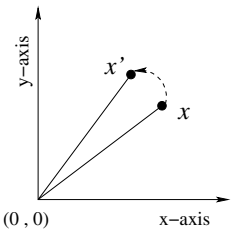


Fig. 3. Rotation in the 2-dimensional space

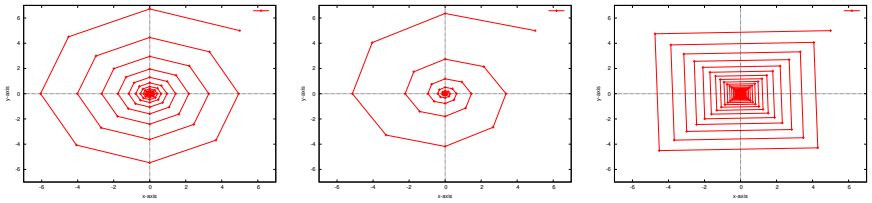
where $0 < \theta < 2\pi$, $0 < r < 1$, and we call $S_2(r,\theta)$ the stable matrix.

As the spiral model given in Eq. (3) shows, while the point x is rotated about the origin (0, 0) by an angle θ, it will be moved closer to the origin (0, 0) by a factor of $1-r$, for r is a value in the open interval (0, 1), i.e., less than one. To make it possible to rotate a point x about an arbitrary point x^*, Eq. (3) has to be modified, as follows:

$$x(k+1) = S_2(r,\theta)\, x(k) - (S_2(r,\theta) - I_2)\, x^*, \tag{4}$$

where again $0 < \theta < 2\pi$, $0 < r < 1$, and I_2 is an identity matrix.

Fig. 4 gives three different search strategies (trajectory of three different spiral models), by using different rotation angle θ and convergence rate r.



(a) $r = 0.95,\ \theta = \pi/4$ (b) $r = 0.90,\ \theta = \pi/4$ (c) $r = 0.95,\ \theta = \pi/2$

Fig. 4. Different spiral models in the two-dimensional space

3 The Proposed Algorithm

3.1 The Concept

Since the search strategy of SO is similar to that of hill-climbing, i.e., to keep updating the best solution (the best center in this case), it lacks mechanisms to disturb the search directions. As such, the center of the spiral may easily get stuck on the local minimum during the convergence process of SO. To solve this problem, a novel spiral optimization is presented in this paper, by adding an oscillation operator to perturb the current search directions (points or individuals) of SO so that it is capable of "escaping the local minimum". In brief, the oscillation operator plays the role of the mutation operator in the genetic algorithm. To increase "intensification" of the

proposed algorithm, one-iteration *k*-means is used to fine-tune the solutions that the proposed algorithm found. Also, to increase "diversification" of the proposed algorithm, the population is divided into several subpopulations, just like the island model of GA, which will then exchange the information they have to each other.

3.2 The distributed Spiral Optimization (dSO)

The proposed algorithm dSO is as outlined in Fig. 5. First, the parameters θ and *r* are set, respectively, to a value in the range (0, 2π) and a value in the range (0, 1), as shown in line 2 of Fig. 5. The population is then divided into N_I subpopulations (i.e., islands) the size of which are $N_P = N_I$ where N_P denotes the population size, and the migration interval α is set. After that, each subpopulation is initialized by randomly

```
1  procedure dSO() {
2      Set the parameters θ and r where 0 < θ < 2π, 0 < r < 1;
3      Initialize each subpopulation, by randomly selecting patterns from the data set as the
       centroids;
4      Select the best center x_c^* for each island;
5      for k = 1 to k_max
6          for c = 1 to N_I // island
7              for each point x_i in island c
8                  x_i(k+1) = Rotation(x_i(k), x_c^*, θ, r);
9                  Oscillation();
10                 OneStepKM();
11             end for
12             x_c^* = UpdateCenter();
13         end for
14         if (k % α == 0) Migration();
15     end for
16 }
```

Fig. 5. Outline of the proposed algorithm dSO

selecting patterns from the data set as the cluster centers. Then, the best center x^*_c (i.e., center with the best fitness value) of each island *c* is selected, as shown in line 4. In the main loop, each point is first rotated, then oscillated, and finally the one-iteration *k*-means applied. Note that in line 8, the points are rotated about the center x^*_c of its island *c*, and the rotation operator is as given in Eq. (4). After that, the center x^*_c of each island is updated. Then, if the migration interval α is hit, the islands will exchange the information they have for each other. In this study, the proposed algorithm will mate $N_I/2$ pairs, by randomly choosing two different elitists from the N_I islands for which the single-point crossover (SPC) is applied to switch the genes of the chromosomes of each pair. Then, dSO will check to see if the termination criterion is met. If the termination condition is met, dSO will terminate and return the best-so-far point as the final solution; otherwise, dSO will continue to perform the three operators (rotation, oscillation, and one-iteration *k*-means) and update the center of each island, as shown in lines 5 to 15.

3.3 The Oscillation Operator

As noted previously, the main purpose of adding the oscillation operator to the SO is to reduce the probability of falling into local minimum, just like the mutation operator of the genetic algorithm [9] does. Here is how it is defined and works. Let F_{max} and F_{min} denote, respectively, the maximum and minimum fitness value of points in the population. Also, let F_x denote the fitness value of point x and δ denote a random number chosen uniformly from the range of $-R$ to R where R is defined as

$$R = \begin{cases} \dfrac{F_x - F_{\min}}{F_{\max} - F_{\min}}, & if \quad F_{\max} \neq F_{\min}; \\ 1, & otherwise. \end{cases} \tag{5}$$

Now, each dimension of the point x is perturbed as follows:

$$x^i = \begin{cases} x^i + \delta \times (x^i_{\max} - x^i), & if \quad \delta \geq 0; \\ x^i + \delta \times (x^i - x^i_{\min}), & otherwise; \end{cases} \tag{6}$$

where $x^i_{\max}$ and $x^i_{\min}$ denote, respectively, the maximum (upper bound) and minimum (lower bound) value of the ith dimension of the data set, meaning that the oscillation operator guarantees that each dimension of a point after perturbation is located within the bounds of that dimension.

4 Experimental Results

4.1 Data Sets and Parameter Settings

The empirical analysis was conducted on an ASUS i7 machine with 2.67 GHz i7- 920 CPU and 4GB of memory using Fedora 12 running Linux 2.6.31. Moreover, all the programs are written in C++ and compiled using g++ (GNU C++ compiler). To evaluate the performance of the proposed algorithm (dSO), we compare it with the spiral optimization [8] and the GKA-clustering algorithm [10], by applying them to the clustering problem for the data sets from UCI [11]. The sum of squared errors (SSE) [12] is used to as a measure of fitness. The parameter settings of SO, dSO, and GKA are as shown in Table 1.

4.2 Results

All the clustering algorithms compared are carried out for 30 runs. The results are as shown in Table 2. Each entry is given as the average plus/minus the standard deviation of 30 runs. For instance, in the case of SO for iris, the result is 78.948 ± 0.003.

As depicted in Table 2, GKA provides a better result than SO. According to our observation, this is so for two reasons: (1) For SO, the points only interact with the

Table 1. Parameter settings of SO, dSO, and GKA

parameter	SO	dSO	GKA
population size	100	100	100
island number	–	10	–
island size	–	10	–
rotation angle (θ)	$\pi/2$	$\pi/2$	–
convergence rate (r)	0.995	0.995	–
oscillation rate	–	0.001	0.001
crossover rate	–	1.0	1.0
migration interval (α)	–	10	–

Table 2. The numerical experiment of SO, dSO, and GKA

Data Set	SO	GKA	dSO
iris	78.948 ± 0.003	**78.941 ± 0.000**	**78.941 ± 0.000**
glass	22.087 ± 0.853	18.288 ± 0.178	**18.248 ± 0.025**
wine	50.912 ± 0.982	**48.954 ± 0.000**	**48.954 ± 0.000**
abalone	30.816 ± 0.876	21.364 ± 0.173	**21.291 ± 0.205**
yeast	46.769 ± 1.075	46.026 ± 0.322	**45.245 ± 0.002**
SPECT.train	233.909 ± 0.465	**233.578 ± 0.000**	**233.578 ± 0.000**

best center and (2) SO lacks a mechanism to perturb the search directions; thus, for SO, the search directions are sensitive to the initial solutions. On the other hand, the results show that dSO matches or outperforms GKA.

The main difference of the three algorithms is that dSO has a higher diversity so that dSO has a higher chance to obtain a better solution during the search. dSO divides the population into several subpopulations so that it has a better chance to search for many potential regions, thus increasing its "diversification." On the other hand, with the addition of the oscillation operator, dSO is able to perturb its search directions, thus increasing its "intensification."

On one hand, in case that dSO provides the same or better results than the other two clustering algorithms, the standard deviation of dSO is very close to zero, meaning that dSO is more robust than the other two clustering algorithms. In other words, the results show that the proposed algorithm is less sensitive to the initial seeds than SO and GKA. On the other hand, it is worth mentioning that although the standard deviation of dSO on the abalone data set is larger than that of GKA; the average result says that dSO has a higher potential to get a better result than SO and GKA.

5 Conclusions

This study is motivated by the observation that the original spiral optimization may easily fall into local minimum during the convergence process. To inherit the features of the spiral optimization, it is easy to implement and only use the rotation operator with the rotation angle θ and the convergence rate *r* to search for the solution space. The proposed algorithm added two operators to the original spiral optimization: the

one-iteration *k*-means and oscillation operators to fine-tune the clustering results and to avoid the premature convergence of the proposed algorithm. Also, dividing the population into several subpopulations eventually increases the search diversity of dSO. In the future, our focus will be on finding more effective rotation and convergence methods to either improve the quality of the clustering results or to apply the proposed algorithm to different optimization problems.

Acknowledgments. This work was supported in part by the National Science Council of Taiwan, ROC, under Contracts NSC101-2221-E-041-012 and NSC99-2221-E-110-052.

References

1. Welch: Algorithmic complexity: Three NP-hard problems in computational statistics. Computation and Simulation (1982)
2. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273 (2003)
3. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face recognition: A literature survey. ACM Computing Surveys, 399–458 (2000)
4. Watada, J., Yamashiro, K.: A data mining approach to consumer behavior. In: Proceedings of the International Conference on Innovative Computing, Information and Control, pp. 652–655 (2006)
5. Geem, Z.W., Kim, J.H., Loganathan, G.V.: A new heuristic optimization algorithm: Harmony search. Simulation, 60–68 (2001)
6. Karaboga, D.: An idea based on Honey Bee Swarm for Numerical Optimization. Erciyes University, Tech. Rep. (2005)
7. Yang, X.-S.: Firefly algorithms for multimodal optimization. In: Watanabe, O., Zeugmann, T. (eds.) SAGA 2009. LNCS, vol. 5792, pp. 169–178. Springer, Heidelberg (2009)
8. Tamura, K., Yasuda, K.: Spiral multipoint search for global optimization. In: International Conference on Machine Learning and Applications, vol. 1, pp. 470–475 (2011)
9. Bandyopadhyay, S., Maulik, U.: An evolutionary technique based on k-means algorithm for optimal clustering in rn. Information Sciences, 221–237 (2002)
10. Krishna, K., Narasimha Murty, M.: Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics 29, 433–439 (1999)
11. UCI-machine learning repository, `http://archive.ics.uci.edu/ml/`
12. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 236–244 (1963)

Recent Development of Metaheuristics for Clustering

Chun-Wei Tsai[1], Wei-Cheng Huang[2], and Ming-Chao Chiang[2]

[1] Department of Applied Informatics and Multimedia,
Chia Nan University of Pharmacy & Science, Tainan, Taiwan
cwtsai0807@gmail.com
[2] Department of Computer Science and Engineering,
National Sun Yat-sen University, Kaohsiung, Taiwan
s96413002@gmail.com, mcchiang@cse.nsysu.edu.tw

Abstract. Metaheuristics have been successfully applied to quite a lot of services, systems, and products frequently found in our daily life. Until now, none of the metaheuristics ever proposed are perfect for all the optimization problems; rather, each algorithm has its pros and cons. Although several high-performance metaheuristics exist, there is still plenty of room to improve the final result they produce and the computation time they take. Since 2001, quite a few number of novel metaheuristics have been developed to provide a better way for solving the optimization problems. A brief review for eight of these novel metaheuristics is given in this paper. To evaluate the performance of these algorithms, we apply them to a well-known combinatorial optimization problem, data clustering, and the results are analyzed and discussed.

Keywords: Metaheuristics, clustering, combinatorial optimization problem.

1 Introduction

The feature of metaheuristics with which we are most familiar is that it is capable of using the rule of thumb to find an approximate optimal solution for an optimization problem in polynomial time [1,2]. Since many optimization problems are either NP-hard or NP-complete [2,3,4], they usually take an unreasonable computation time to find a reasonable solution by using an exhaustive search algorithm. To solve this problem, metaheuristics use an iterative search strategy to find an approximate optimal solution by using a limited computation resource, such as computing power and computation time.

Clustering [5,6,7] is indeed one of the most well-known optimization techniques which have been applied to many problem domains, such as search engine [8], document analysis [9], biological information [10], and face recognition[11]. Generally speaking, the main feature of clustering [12,13] is that no a priori knowledge is required to analyze the input data (patterns). The clustering problem is generally defined as: Given a set of n patterns $X = \{x_1, x_2, ..., x_n\}$ in d-dimensional space, partition the set X into k clusters $\prod = \{\pi_1, \pi_2, ..., \pi_k\}$ that

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_93,

minimize some predefined criteria (e.g., sum of squared error (SSE), entropy, f-measure, or accuracy). Although the traditional deterministic algorithms (e.g., k-mean) are simple and easy to implement, the results are extremely sensitive to the initial solution, a problem that metaheuristics do not have. As such, we can envision that metaheuristics will continue to play an important role in solving the clustering problem [14,15].

The remainder of the paper is organized as follows. After a brief introduction to metaheuristics in Section 1, Section 2 proceeds to describe eight novel metaheuristics that were developed during the decade 2001 to 2011. Then, these metaheuristics for clustering are compared and analyzed in Section 3. Finally, changes and potentials of these new search algorithms themselves, open issues, and future trends are presented in Section 4.

2 The New Metaheuristic Algorithms

Table 1 shows the eight kinds of metaheuristics with different design inspirations to be described in this section. To make it easier to differentiate these algorithms, the full name of each algorithm, the acronym, the behavior, and the year it was published are also given in this table.

Table 1. Eight novel metaheuristic algorithms

Algorithm	Acronym	Behavior	Year
Artificial bee colony	ABC	insects	2005 [16]
Bat algorithm	BA[a]	animals	2010 [17]
Bioluminescent swarm optimization	BSO	insects	2011 [18]
Firefly algorithm	FA	animals	2009 [19]
Gravitational search algorithm	GSA	natural	2009 [20]
Honey bee mating	HBM	insects	2001 [21]
Harmony search	HS	human	2001 [22]
Imperialist competitive algorithm	ICA	human	2007 [23]

[a] In this paper, we will use BA instead of BAT to denote the bat algorithm so that all the algorithms described herein go by the name of their acronyms.

The artificial bee colony (ABC) was designed to emulate the behavior of bees during foraging. It was first proposed by Karaboga [16] in 2005 and used to solve the continuous function optimization problem. ABC has also been used to solve combinatorial optimization problems, such as scheduling [24] and clustering [25,26]. In [27], Akay and Karaboga pointed out that the scout limitation parameter setting will affect the final results of ABC. A later study [28] also showed that ABC is not very sensitive to initialization. Presented by Yang [17] in 2010, the bat algorithm (BA) was inspired by the bat echolocation behavior. Its applications can be seen on several different research domains, such as clustering [29], engineering optimization [30], and multi-objective optimization [31]. To enhance the performance of BA, Khan

et al. [29] combined BA with the fuzzy concept to solve the clustering problem. In a later study [30], Yang and Gandomi pointed out that the convergence rate and the result of BA will be affected by the setting of the loudness A_i and pulse emission rate r_i of bats [32]. Proposed by de Oliveira et al. [18] in 2011, the bioluminescent swarm optimization (BSO) was designed to emulate the behavior of glowworm fly. The main idea came from the glowworm swarm optimization (GSO), proposed by Krishnanand and Ghose [33] in 2005. Krishnanand and Ghose [34] pointed out that the max sensing light of agent in GSO is hard to limit, but it may be useful to apply it to multiple solutions in the high-dimensional space.

Another interesting metaheuristics is proposed by Yang [19] called the firefly algorithm (FA) which is inspired by the characteristics of fireflies being attracted by the brightness of the others. It was _rst used in solving the function optimization problem, and the performance is quite outstanding [35,36]. In [37,35], Levy flights was combined with FA to improve its capability to perturb the solutions locally. In a later study [38], Giannakouris et al. attempted to combine FA with ACO to leverage the advantages of each. An alternative way was presented in [39] which uses refined initial solutions to improve the quality of the results. The gravitational search algorithm (GSA) is designed to emulate the behavior of nature gravitation. It was first proposed by Rashedi et al. [20] in 2009 and used to solve the function optimization problem [40,41]. To enhance the search performance of GSA, Sarafrazi et al. [40] used the disruption operator to prevent two agents from getting too close to each other. In a later study [42], Li and Zhou improved the GSA by adding the best agent information to update the position. Besides, in [41], Askari and Zahiri also used the fuzzy concept to determine the range of agent's movement.

The honey bee mating (HBM) is inspired by queen bee mating. It was proposed by Abbass [21] in 2001. In [43], Marinakis et al. combined the greedy randomized adaptive search with HBM to solve the clustering problem. A later study [44] used the fuzzy set and HBM to solve the multi-objective distribution feeder reconfiguration problem. In [45], Chang also used the honey bees policy iteration to solve the stochastic dynamic programming problem. In [46], Marinakis et al. pointed out that using more than one queen can obtain a better quality. The harmony search (HS) is inspired by musicians creating harmony. It was first proposed by Geem et al. [22] in 2001 and used to solve the traveling salesman problem, the specific academic optimization problem, and the least cost pipe network problem. In the survey of HS [47], Geem pointed out that HS performs outstandingly for the problems to which it applies. In [48], Fesanghary et al. used sequential quadratic programming to speed up the convergence rate of HS. In [49], Li and Li combined HS with PSO for which HS plays the role of fine-tuning the solution found by PSO. In a later study [50], Wang and Huang pointed out that a better initial harmony memory will improve its performance. In [51], Omran and Mahdavi used the best harmony as a tuning variable to create new harmony. Besides, the studies given in [52,53,54] focus on adjusting the harmony memory considering rate (HMCR) and the pitch adjusting rate (PAR). The imperialist competitive algorithm (ICA) is inspired by imperialist competitive. It was _rst proposed by Atashpaz-Gargari and Lucas [23] in 2007 and used to solve the function

optimization problem. The studies given in [55,56,57] all focus on modifying the moving angle factor of assimilation policy, such as in [58], Abdechiri et al. used a Gaussian probability model to adjust the angle. In a later study [59], Zhang et al. used a linear perturbation model to adjust the position of colony after the assimilation policy.

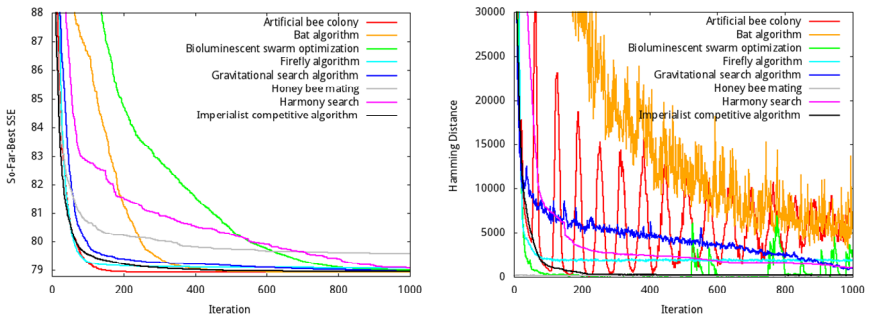
3 Experimental Results

3.1 Parameter Settings and Data Sets

The empirical analysis was conducted on an ASUS i7 PC with 2.67 GHz i7-920 CPU and 4GB of memory running Fedora 12 with Linux 2.6.31. Moreover, all the programs are written in C++ and compiled using g++ (GNU C++ compiler). In addition, the IRIS from UCI [60] is used to compare the performance of the metaheuristics in this study. For the simulation, each algorithm is carried out for 30 runs, and the maximum number of iterations for each run is set equal to 1,000.

3.2 Simulation Results

As shown in Fig. 1(a), the simulation results show that for IRIS, all these metaheuristics provide results that are close to each other. Fig. 1(a) shows not only the quality but also the convergence speed of each algorithm. Fig. 1(b) shows the hamming distances of each algorithm during the evolutionary process. All the solutions are integer encoded and are of length 150; all the populations are of size 50, implying that if all the solutions are different, the hamming distance will be maximized, up to $[50 \times 150 \times (50-1)]/2 = 183{,}750$.



(a) The so-far-best SSE (b) The hamming distance

Fig. 1. The results of applying all the eight metaheuristics to IRIS

As shown in Fig. 1(b), for ABC, the scout bee operator is used to move the solutions around in the search space periodically and randomly while at the same time the onlooker bee operator is used as the selection mechanism which will decrease the

diversity of the population. This explains why the hamming distances for ABC oscillate cyclically. BA uses loudness and pluse emission rate to control the quality and diversity. Although BSO converges, when it uses the mass extinction strategy, the diversity may be increased at the later stages during the evolution process. FA uses a random perturbation strategy for each firefly agent to maintain the diversity of the population. GSA uses the gravitation and the number of attracted agents to dynamically control the diversity. HBM can be regarded as a single-solution-based algorithm, which evolves around the queen to generate children with a low diversity. For HS, the memory of musicians will greatly affect the diversity of the solution. ICA enhances the diversity of the solution by assigning colonies to different empires, but imperialistic competition will eventually move all the colonies to the same empire at later iteration.

4 Conclusions and Future Work

This paper briey reviews the most recent development in the field of metaheuristics. These novel algorithms were inspired from the behavior of insects, animals, nature, and human but are not paid particular attention until now. All metaheuristics have advantages and disadvantages. The ground truth is that there simply does not exist any metaheuristics that can solve all the combinatorial optimization problems. Therefore, how to efficiently apply metaheuristics to large and complex problems and find an acceptable solution in a reasonable time are two of the most critical research issues in the area of metaheuristics. In this paper, these new and relatively young metaheuristics are applied to the clustering problem. The quality of the results and the diversity are analyzed and discussed. The above analysis helps us better understand the characteristics of these algorithms. In the future, we will attempt to present a systematic description of these algorithms and a more detailed discussion about how to apply them to other problem domains.

References

1. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers & Operations Research 13(5), 533–549 (1986)
2. Blum, C., Roli, A.: Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys 35(3), 268–308 (2003)
3. William, Welch, J.: Algorithmic complexity: Three np-hard problems in computational statistics. Journal of Statistical Computation and Simulation 15(1), 17–25 (1982)
4. Garey, M.R., Johnson, D.S.: Computers and Intractability; A Guide to the Theory of NP-Completeness. WH Freeman and Company, New York (1990)
5. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)
6. Xu, R., Wunsch II, D.: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005)
7. Rai, P., Singh, S.: A survey of clustering techniques. International Journal of Computer Applications 7(12), 156–162 (2010)

8. Carpineto, C., Osiński, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Computing Surveys 41(3), 17:1–17:38 (2009)
9. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of Conference on Research and Development in Information Retrieval, New York, NY, USA, pp. 267–273 (2003)
10. Getz, G., Gal, H., Kela, I., Notterman, D.A., Domany, E.: Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics 19(9), 1079–1089 (2003)
11. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2003)
12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)
13. Xu, R., Wunsch II, D.: Survey of clustering algorithms. IEEE Transaction on Neural Networks 16(3), 645–678 (2005)
14. Bianchi, L., Dorigo, M., Gambardella, L., Gutjahr, W.: A survey on metaheuristics for tochastic combinatorial optimization. Natural Computing 8, 239–287 (2009)
15. Puchinger, J., Raidl, G.R.: Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2005. LNCS, vol. 3562, pp. 41–53. Springer, Heidelberg (2005)
16. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Technical Report, Erciyes University, Engineering Faculty, Computer Engineering (2005)
17. Yang, X.-S.: A new metaheuristic bat-inspired algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)
18. de Oliveira, D.R., Parpinelli, R.S., Lopes, H.S.: Bioluminescent Swarm Optimization Algorithm. Evolutionary Algorithms (2011)
19. Yang, X.-S.: Firefly algorithms for multimodal optimization. In: Watanabe, O., Zeugmann, T. (eds.) SAGA 2009. LNCS, vol. 5792, pp. 169–178. Springer, Heidelberg (2009)
20. Rashedi, E., Nezamabadi-pour, H., Saryazdi, S.: GSA: A Gravitational Search Algorithm. Information Sciences 179(13), 2232–2248 (2009)
21. Abbass, H.: MBO: marriage in honey bees optimization-a Haplometrosis polygynous swarming approach. In: Proceedings of Computation Congress on Evolutionary Computation, vol. 1, pp. 207–214 (2001)
22. Geem, Z.W., Kim, J.H., Loganathan, G.: A new heuristic optimization algorithm: Harmony search. Simulation 76(2), 60–68 (2001)
23. Atashpaz-Gargari, E., Lucas, C.: Imperialist competitive algorithm: An algrorithm for optimization inspired by imperialistic competition. In: Proceedings of the IEEE Congress on Evolutionary Computation, pp. 4661–4667 (2007)
24. Pan, Q.-K., Tasgetiren, M.F., Suganthan, P., Chua, T.: A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Information Sciences 181(12), 2455–2468 (2011)
25. Karaboga, D., Ozturk, C.: A novel clustering approach: artificial Bee Colony (ABC) algorithm. Applied Soft Computing 11(1), 652–657 (2011)
26. Zhang, Y., Wu, L., Wang, S., Huo, Y.: Chaotic artificial bee colony used for cluster analysis. Intelligent Computing and Information Science 134, 205–211 (2011)
27. Akay, B., Karaboga, D.: Parameter tuning for the artificial bee colony algorithm. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 608–619. Springer, Heidelberg (2009)

28. Akay, B., Karaboga, D.: A modified artificial bee colony algorithm for real parameter optimization. Information Sciences 192, 120–142 (2012)
29. Khan, K., Nikov, A., Sahai, A.: A fuzzy bat clustering method for ergonomic screening of office workplaces. In: Dicheva, D., Markov, Z., Stefanova, E. (eds.) Software, Services and Semantic Technologies S3T 2011. AISC, vol. 101, pp. 59–66. Springer, Heidelberg (2011)
30. Yang, X.-S., Gandomi, A.H.: Bat algorithm: a novel approach for global engineering optimization. Engineering Computations 29(5), 464–483 (2012)
31. Yang, X.-S.: Bat algorithm for multi-objective optimization. International Journal of Bio-Inspired Computation 3(5), 4267–4274 (2011)
32. Damodaram, R., Valarmathi, M.L.: Phishing website detection and optimization using modified bat algorithm. International Journal of Engineering Research and Applications 2(1), 870–876 (2012)
33. Krishnanand, K., Ghose, D.: Detection of multiple source locations using a glowworm metaphor with applications to collective robotics. In: Proceedings of Computation Congress on Swarm Intelligence Symposium, pp. 84–91 (2005)
34. Krishnanand, K., Ghose, D.: Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. Swarm Intelligence 3, 87–124 (2009)
35. Yang, X.-S., Deb, S.: Eagle strategy using lévy walk and firefly algorithms for stochastic optimization. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 101–111. Springer, Heidelberg (2010)
36. Gandomi, A., Yang, X.-S., Talatahari, S., Alavi, A.: Firey algorithm with chaos. Communications in Nonlinear Science and Numerical Simulation (2012)
37. Yang, X.-S.: Firey algorithm, lévy ights and global optimization. In: Research and Development in Intelligent Systems, pp. 209–218 (2010)
38. Giannakouris, G., Vassiliadis, V., Dounias, G.: Experimental study on a hybrid nature-inspired algorithm for financial portfolio optimization. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 101–111. Springer, Heidelberg (2010)
39. Łukasik, S., Żak, S.: Firefly algorithm for continuous constrained optimization tasks. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 97–106. Springer, Heidelberg (2009)
40. Sarafrazi, S., Nezamabadi-pour, H., Saryazdi, S.: Disruption: A new operator in gravitational search algorithm. Scientia Iranica 18(3), 539–548 (2011)
41. Askari, H., Zahiri, S.-H.: Decision function estimation using intelligent gravitational search algorithm. International Journal of Machine Learning and Cybernetics 3, 163–172 (2012)
42. Li, C., Zhou, J.: Parameters identification of hydraulic turbine governing system using improved gravitational search algorithm. Energy Conversion and Management 52(1), 374–381 (2011)
43. Marinakis, Y., Marinaki, M., Matsatsinis, N.F.: A hybrid clustering algorithm based on honey bees mating optimization and greedy randomized adaptive search procedure. In: Maniezzo, V., Battiti, R., Watson, J.-P. (eds.) LION 2007 II. LNCS, vol. 5313, pp. 138–152. Springer, Heidelberg (2008)
44. Niknam, T.: Application of honey-bee mating optimization on state estimation of a power distribution system including distributed generators. Journal of Zhejiang University - Science A 9, 1753–1764 (2008)
45. Chang, H.: Converging marriage in honey-bees optimization and application to stochastic dynamic programming. Journal of Global Optimization 35, 423–441 (2006)

46. Marinakis, Y., Marinaki, M., Dounias, G.: Honey bees mating optimization algorithm for the euclidean traveling salesman problem. Information Sciences 181(20), 4684–4698 (2011)
47. Geem, Z.W.: Music-Inspired Harmony Search Algorithm: Theory and Applications. Springer (2009)
48. Fesanghary, M., Mahdavi, M., Minary-Jolandan, M., Alizadeh, Y.: Hybridizing harmony search algorithm with sequential quadratic programming for engineering optimization problems. In: Computer Methods in Applied Mechanics and Engineering, vol. 197(3340), pp. 3080–3091 (2008)
49. Qi Li, H., Li, L.: A novel hybrid particle swarm optimization algorithm combined with harmony search for high dimensional optimization problems. In: International Conference on Intelligent Pervasive Computing, pp. 94–97 (2007)
50. Wang, C.-M., Huang, Y.-F.: Self-adaptive harmony search algorithm for optimization. Expert Systems with Applications 37(4), 2826–2837 (2010)
51. Omran, M.G., Mahdavi, M.: Global-best harmony search. Applied Mathematics and Computation 198(2), 643–656 (2008)
52. Jaberipour, M., Khorram, E.: Two improved harmony search algorithms for solving engineering optimization problems. Communications in Nonlinear Science and Numerical Simulation 15(11), 3316–3331 (2010)
53. Pan, Q.-K., Suganthan, P., Tasgetiren, M.F., Liang, J.: A self-adaptive global best harmony search algorithm for continuous optimization problems. Applied Mathematics and Computation 216(3), 830–848 (2010)
54. Al-Betar, M.A., Khader, A.T., Liao, I.Y.: A harmony search with multi-pitch adjusting rate for the university course timetabling. In: Geem, Z.W. (ed.) Recent Advances In Harmony Search Algorithm. SCI, vol. 270, pp. 147–161. Springer, Heidelberg (2010)
55. Abdechiri, M., Faez, K., Bahrami, H.: Neural network learning based on chaotic imperialist competitive algorithm. In: Proceedings of the International Workshop on Intelligent Systems and Applications, pp. 1–5 (2010)
56. Duan, H., Xu, C., Liu, S., Shao, S.: Template matching using chaotic imperialist competitive algorithm. Pattern Recognition Letters 31(13), 1868–1875 (2010)
57. Talatahari, S., Azar, B.F., Sheikholeslami, R., Gandomi, A.: Imperialist competitive algorithm combined with chaos for global optimization. Communications in Nonlinear Science and Numerical Simulation 17(3), 1312–1319 (2012)
58. Abdechiri, M., Faez, K., Bahrami, H.: Adaptive imperialist competitive algorithm (AICA). In: Proceedings of the International Conference on Cognitive Informatics, pp. 940–945 (2010)
59. Zhang, Y., Wang, Y., Peng, C.: Improved imperialist competitive algorithm for constrained optimization. In: Proceedings of the International Forum on Computer Science-Technology and Applications, vol. 1, pp. 204–207 (2009)
60. UCI-machine learning repository, `http://archive.ics.uci.edu/ml/`

The Originality of a Leader for Cooperative Learning

Po-Jen Chuang[1] and Chu-Sing Yang[2]

[1] National Sun Yat-sen University
Department of Computer Science and Engineering
Kaohsiung, Taiwan, R.O.C.
[2] National Cheng Kung University
Department of Electrical Engineering
Tainan, Taiwan, R.O.C.
flypojen@gmail.com

Abstract. A team work begins with cooperative learning, the responsibility of associators, abilities of individuals, and team progress are rewarded for helping and sharing to each other. The leader leads the group comprehending tasks, directing discussion, and progressing in studying. For better cooperation, we explode the previous research of the grouping strategy - pairing strategy [1] which enhances the learning and testing results of students based on the relationship of social network. Basing on the society relationships, this method provides members of the groups to learn from or mimic learners who have good relationships with them. This paper discusses about supports, including parents, teachers, and members of the group, for a leader leading the group to a higher achievement. The specific factors are also discussed, that the parents' support is the biggest among the others.

Keywords: Social Learning, Cooperative Learning, Grouping Strategy, Learning Strategy, Learning Achievement, Regression Line.

1 Introduction

Traditional learning methods have been gradually shifted from individual learning to cooperative learning, because of the ubiquity of e-Learning. By using the fast interaction between people and the obtainment of the tests, the learning method adopts not only the grades, but the selection of partner is also important [2]. Cooperative learning begins with groups. A good grouping method has to take into account many factors, such as learning achievement, the depth of teaching materials and categories. As such, personal traits, good relationships, learning behaviors all have something to with the learning results. For the same reason, the friends of an individual can be considered as the incentive of cooperative learning [3]. The pair cooperation takes place the learning of cooperation and the other grouping strategy can base on it. The leader can be interpreted as one who has some qualifications, leads people toward the goal and is responsible for it. Bennis [4] pointed out that a leader should have certain behavioral tendencies, such as the drive for development, willingness to take up challenges, innovation, having long-term goals, creativity, and

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_94,

always thinking about when to do it or how to achieve it. Tead [5] thinks leadership is an activity which works people went to the pursuit of the goals. Stogdill [6], [7] pointed that leadership leads people to set a goal and accomplish the goal. If it is said that the leadership is a condition that affects the behavior of the crowd, then Hemphill and Coons [8] agree with Stogdill. Lao [9] and Shie [10] pointed out that leadership is a behavior that a leader leads adherents together to accomplish the targets in the interacting process of a group. The leadership can be summarized by Wu [11] that a leader affects adherents to effectively accomplish the goals in a certain circumstances.

The rest of this paer is organized as follows. Section 2 gives the related work about the development of characteristics of a leader. The concepts and details of the proposed algorithm are given in Section 3, Section 4 shows and analyzes the differences among varied relationships. Section 5 concludes the paper and suggests some possible future research directions

2 Related Works

Therefore, we can easily find that the leaders possess certain characteristics including 1) A strong desire for high achievement. Professor Ghiselli [12] surveyed 90 different industries in the US and found that effective leaders tend to be those with high achievement. These people have a strong potential to realize their goals, nally become successful and treat them as the ladder achieving their goals. 2) Spirits of perseverance and persistence. That is the representation of a stronger dedication to persistence, patience, and confidence. 3) Being systemic.

According to the historical backgroound, the theories studied on leadership can be divided into trait theory, behavior theory, situational theory, and new leadership theory [13].

- Before late 1940s, the studies focused on the trait theory which believes that the ability of leadership was born. The leaders are born and demonstrate distinct attributes in some cases during this period.
- From late 1940s to late 1960s, the studies focused on the behavior pattern theory, which emphasised the connection between leadership effectiveness and the leaders' behavior. In this period, the studies still took the personal qualities of a leader as the main consideration, but emphasized the ability of leadership can be achieved through learning for highlighting the importance of the leader's behavior.
- From late 1960s to early 1980s, the studies adopted the approach of contingency situational theory, which suggested that the effectiveness of a leader was driven by a combination of factors rather than simply by the leader's traits. The focus of the studies were extended to the leader, the team members and the context of the situation. It assumes that the leader could motivate the team members within the context of the situation and impact their behavior.
- Before early 1980s, the studies focused on visionary leaders. This theory combined the viewpoints from the three above-mentioned theories, drawing on one's strengths to compensate for the other's shortcomings and constructed a new leadership theory that could cope with the rapid changes of the environment in the future. These

leaders not only possessed the ability to influence their team members, but were also able to find the optimal approach for development based on the organization's strengths and weaknesses. Hersey and Blanchard [14], [15] indicated when selecting a leader, the keys to the managerial grid and situational leadership were his attitude and behavior.

3 Proposed Method

In the experiences of the cooperative learning, students get supports from the partners of their groups, advices from teachers, and emotional and knowledgable concern by their parents. These theories of leader's originality conclude a common goal of figuring out the factors that develop a leader and how to achieve effective leadership. In cooperative learning, the leader of the group also has the same advantages of leading group learning. A leader has a better ability of communication to re-organize the teaching of teachers and provides a better way of understanding for the crews. Also, the leader can translate questions from the crews to teachers. To summarize the task of a leader, the following can be described.

1. Communications and advices from one to the partners in the group.
2. Communication with teachers.
3. Assistance of family.

We provide the definition to decide the leader:

Definition 1. (Leadership) The ability of a leader j is the maximum value of L_j defined as

$$L_j = \mathrm{MAX}_j(R_j + E_j + P_j) \tag{1}$$

where R_j is the relationship of j with partner, E_j is the ability of seeking helps from teachers, $E_j = C_{jt} * A_{jt} + L_{jt}$, communicates with teacher and ask advice from teacher, and like to talk to teacher, and P_j is the level of parents' care time. In this case, we don not concern 'dislike', because that what ever the student likes teacher or not, the teacher needs to like the student. Besides, the teacher needs to concern the student.

4 Experimental Results

The experiment works on the sixth grade of elementary school in Kaohsiung. The participants are 68 students from two classes of sixth grade students, with one class as the social pairs and the other class as the random pairs, and their scores are collected throughout the academic year. We take subject Mandarin as the learning material and test scope. The experiment spans the midterms and finals and the grouping is conducted twice, once in the first semester while the other in the second semester. Students learn in their classes and by the method of cooperative learning STAD, which they learn from the teachers, study together and each accomplishment will be

the group achievement. The student nformation and scores are input by their teachers. Every time, questionnaires are given to rebuild the friendship sets. The experiment uses first exam in the first semester of grade 5 as the pre-test, the final exam in the first semester of grade 6 as the post-test, and friendship sets conducted in totally a year. The data of the first semester of grade 5 is not used because students are not particularly familiar with the grouping experiment. Instead, the data of the second semester of grade 6 is not used because students will ready graduate from elementary.

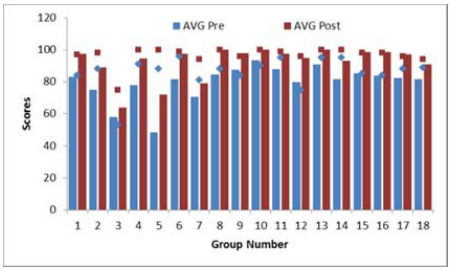


Fig. 1. The achievement relation of the leader and crews

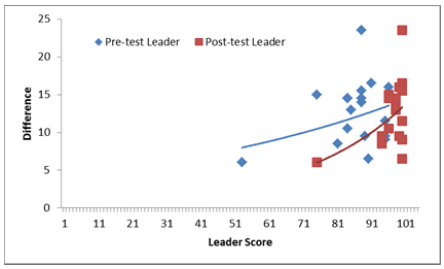


Fig. 2. The difference between pre-test and post-test

In the grouping pair, the appointment of a leader is based on academic score, relationship with members, communication with teachers, and the parents' support. Apparently, a good group achievement will be presented with the members confided their dependence to the leadership style and received the leader. In this section, we verify three statements: 1) whether the leader score will lead the group score? 2) Whether the relationship between the leader and teachers will affect the group score? 3) Whether the family support of the leader could affect the group score?

Fig. 1 shows the pre-test and post-test scores of the group average as well as the leader score of the experimental group. The blue color indicates the pre-test, the bar indicates the group average and the diamond indicates the leader score. The red color indicates the post-test and the square indicates the leader score. The total eighteen groups are because of the pair strategy, which implies two students share the same leader with other student.

As the figure shows:

- The post achievements of the groups are increased. This result indicates the practice of group learning.
- The leaders' score are promoted which implies the group learning can also help the student of higher achievement.

However, this result leads to a question, does the higher difference between pre-test and post-test implies the higher achievement of the leader. Fig. 2 shows the difference of two tests, where x-axis shows the leader score and y-axis shows the difference of pre-test and post-test. The red square indicates the post-test score of the leader and the blue diamond indicates the pre-test. The red and blue curves show the trend of the

data, which indicates the leader score of both pre-test and post-test implies the growth of group scores.

As the Fig. 2 shows, the higher leader score can really promote the higher difference, both in the pre-test and post-test. The leader connects the teacher and the group learning. A good relation with the teacher can also help the group learning in a clear way. In contrast, the rare connection between the leader and the teacher can probably pull down the effciency of learning.

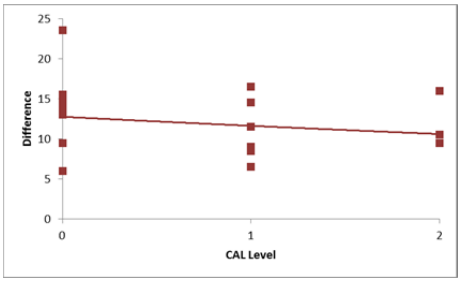


Fig. 3. The difference of the relation between the leader and the teacher

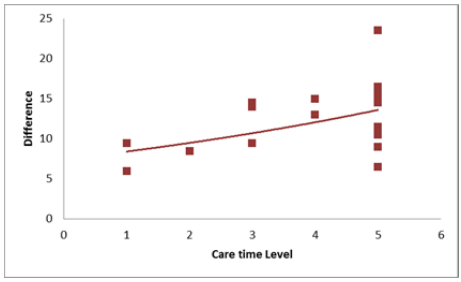


Fig. 4. The difference of the relation between the leader and the parents

Fig. 3 shows the value of E_i of the leader relates the difference of the pre-test and the post-test. The *x*-axis shows the CAL value, which are the organization of communication, advice, and like, and the *y*-axis shows the difference. The CAL values have only three kinds and red curve shows the trend that the differences are equally distributed in these three values. This result shows that the connection between the leader and the teacher is not the main option to affect the group learning.

As the main option of the leader pick, the parent's care time can help the leader learning. Moreover, the leader can be confident to lead the group learning. Fig. 4 shows the relation between the parents' care time of the leader and the difference. The values of parents' care time have five values and the higher value indicates the more care. The red curve shows the trend that the more care time, the higher difference. Obviously, the support of the family is powerful for learning.

5 Conclusion

Cooperative learning opens a way to achieve group goals by learning from team members. In the meantime, the leader takes own group to comprehend the achievements. We examined the supports of a leader, relationship between the leader and crews, relation with teachers, and care from the parent, and found that the group leader who obtains family support can demonstrate a higher level of effectiveness in group learning.

Based on our research, grouping of more than three people can utilize our proposed strategy and achieve the purpose of cooperative learning. In the future, we intend to include discussions on grouping of multiple people (three or four) and compare the

results with grouping in pairs [16]. In the meantime, we expect to explore further on the leadership effectiveness of the group head, measuring whether the leader of a group with more people are more efficient than that with a pair. Qualitative research deeply recognizes the behavior of human being and reasons of methods, not only the operation of research [17]. Cooperative learning connects to the learning styles and communication ways of students. Qualitative research will help us discussing data collection, analyzing teaching data, feedback, and validation, reliability, and ethics among members [18], [19].

References

1. Chuang, P.J., Chiang, M.C., Yang, C.S., Tsai, C.W.: Social networks-based adaptive pairing strategy for cooperative learning. Educational Technology & Society 15(3), 226–239 (2012)
2. Rosenberg, M.: E-learning: Strategies for Delivering Knowledge in the Digital Age. McGraw-Hill Professional (2001)
3. Mason, R.: E-learning and social networking handbook: Resources for higher education. Routledge (2008)
4. Bennis, W.: Leadership theory and administrative behavior: The problem of authority. Administrative Science Quarterly, 259–301 (1959)
5. Tead, O.: The art of leadership (1935)
6. Stogdill, R.: Leadership, membership and organization. Psychological Bulletin 47(1), 1 (1950)
7. Stogdill, R.: Handbook of leadership: A survey of theory and research. Free Press (1974)
8. Hemphill, J., Coons, A.: Development of the leader behavior description questionnaire. Leader Behavior: Its Description and Measurement 6, 38 (1957)
9. Lee, Y.-T.: The Study of Leadership Theory
10. Hsieh, W.C.: Educational Administration: Theories and Practice. Win-Join Book Co., Taipei (1987)
11. Wu, D., Chang, J.S., Chen, D.Y., Lai, W.Y., et al.: Public Administration (1). National Open University, Taipei (1994)
12. Ghiselli, E.: Explorations in managerial talent. Goodyear Pub. Co. (1971)
13. Bryman, A.: Charisma and leadership in organizations. Sage, London (1992)
14. Hersey, P., Blanchard, K.: Leadership style: Attitudes and behaviors. American Society for Training & Development (1982)
15. Hersey, P., Blanchard, K., Johnson, D.: Management of organizational behavior. Prentice-Hall, Englewood Cliffs (1988)
16. Marie Rysavy, S., Sales, G.: Cooperative learning in computer-based instruction. Educational Technology Research and Development 39, 70–79 (1991)
17. Denzin, N., Lincoln, Y.: Handbook of qualitative research. Sage Publications, Inc. (1994)
18. Bogdan, R., Biklen, S.: Qualitative research in education. An introduction totheory and methods. ERIC (1998)
19. Merriam, S.B.: Qualitative Research and Case Study Applications in Education.Revised and Expanded from Case Study Research in Education. ERIC (1998)

A Hybrid Ant-Bee Colony Optimization for Solving Traveling Salesman Problem with Competitive Agents

Abba Suganda Girsang[1], Chun-Wei Tsai[2], and Chu-Sing Yang[1]

[1] Inst. of Computer and Communication Engineering, Dept. of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, ROC
[2] Department of Information Technology, Chia Nan University of Pharmacy Science, Tainan, Taiwan, ROC
gandagirsang@yahoo.com, cwtsai0807@gmail.com, csyang@ee.ncku.edu.tw

Abstract. This paper presents a new method called hybrid ant bee colony optimization (HABCO) for solving traveling salesman problem which combines ant colony system (ACS), bee colony optimization (BCO) and ELU-Ants. The agents, called ant-bees, are grouped into three types, scout, follower, recruiter at each stages as BCO algorithm. However, constructing tours such as choosing nodes, and updating pheromone are built by ACS method. To evaluate the performance of the proposed algorithm, HABCO is performed on several benchmark datasets and compared to ACS and BCO. The experimental results show that HABCO achieves the better solution, either with or without 2opt.

Keywords: Hybrid, Ant Colony System, Bee Colony System, Traveling Salesman Problem.

1 Introduction

In the current era, intelligent algorithms inspired by nature behaviors play the role of solving several optimization problems. Ant colony optimization (ACO), particle swarm optimization (PSO), and bee colony optimization (BCO) are the samples being categorized for those intelligent algorithms[1-14].

Ant colony optimization (ACO) was firstly developed by Dorigo, who introduced ant system (AS) to solve traveling salesman problem [1]. Ants tend to choose the path with the strongest pheromone to find the shortest tour for the food. Dorigo then extended this algorithm by introduce algorithm ant colony system (ACS) [2]. Since then, many researchers also continued this approach for better results on various optimization problems [3-6]. Max-min ant system (MMAS) was proposed by Stutzle and Hoos to solve TSP [3]. They proposed the pheromone level update that is bounded the pheromone level between τ_{max} and τ_{min}. Naimi and Taherinejad proposed ELU-Ants and KCC-Ants to modify ACS by introduce new local update [4]. This research emphasized that the update pheromone on earlier step is more than the later step. Chen and Chien solved TSP problem by combining genetic algorithm (GA) such as selection, crossover, mutation and ACS with communication strategy [5].

James J. (Jong Hyuk) Park et al. (eds.), *Mobile, Ubiquitous, and Intelligent Computing*, Lecture Notes in Electrical Engineering 274,
DOI: 10.1007/978-3-642-40675-1_95,

They also proposed algorithm to solve TSP by combining GA, simulated annealing, ACS and particle swarm optimization (PSO) [6]. Besides solving TSP, ACO are also used to solve other combinational optimization problem such as job-shop scheduling problem [7], quadratic assignment problem [8] and so forth.

Another swarm intelligent algorithm bee colony optimization (BCO), was firstly proposed by Lucic et al [9-10] to solve TSP. BCO is designed to create the multi-agent system (colony of artificial bees) for solving the combinatorial optimization problems. Swarm behavior encourages generating the algorithms such as bee system, artificial bee colony (ABC), marriage in honey bee optimization (MBO), etc. Swarm bees are also used for solving vehicle routing problem [11], numerical optimization [12], data mining [13], job shop scheduling [14] and so forth. In BCO, after traveling some nectars or stage, each bees backs to the hive to communicate their way. In the hive, they group into three types and perform waggle dance to identify their quality of food source. The longer dance indicates the better quality of nectars found. Bees exchange information to each other for their next stage-tour. Bees finding bad-source food tend to follow the others with good-source food at next stage.

On ant behavior, once an ant starts visiting a city, it will remain active until the end of the complete visiting all cities. Ants might travel the good choice or bad choice in their earlier steps. Good choice in the earlier step means the tour being relatively short when ants continue to complete any probability remainder tour. Contrary, bad choice may cause successively bad selections in later parts and bad result at the end. So based on discussion above, this research is tried to develop the algorithm that considers the beginning step as the more important step to construct the tour than the last step. Logically, an agent with bad choice of traveling some cities (part of tour) should be discarded but an agent with good choice should be duplicated at the beginning step. The action on this earlier step guarantees the agent will be more competitive.

2 Proposed Algorithm

2.1 The Concept

In this section, a new algorithm, hybrid ant-bee colony optimization (HABCO) for solving the traveling salesman problem is presented. The competitive agents are built by considering the performance on the earlier step in constructing the tour. The tour is split into some parts of tour. This part of tour can be identified as the stage in BCO which consists some cities. The stage-tour agent constructed is expected to judge the quality agent. On BCO, after traveling one stage, bees back to hive and group into three types based on their quality. To travel each cities on one stage, the ant-bee uses the role of ACS.

The judgment of the quality of each agents especially on early stages becomes important issues. The judgment should guarantee the agent whose good stage-tour at beginning step will probably achieve the good complete tour. In ACS there are two important variables to construct the tour, pheromone and distance. After traveling some cities, says one stage, the pheromone and distance cities gathered in one stage are investigated to conclude whether the both variables can be to a judgment of the

quality agent. One benchmark TSP, *st70*, is investigated to see impact of the distance and amount pheromone at each stages. The agent (ant) travels all cities stage by stage with follow the ACS role, and there is no interaction likes BCO in each stages.

In the observation, considering by calculating the distance traveled at one stage is not reliable. The second observation is to rely on the total amount of pheromone in one stage tour traveled by each ants. It shows that the ant which gathered a lot of pheromone at first stage tends to be a good or best ant in the complete tour. The other hand, the ant which get the shortest distance probably gets the high pheromone on its first stage. So, it is more reasonable that total amount of pheromone at each stages indicates the quality sub-tour. The greater total pheromone that the ant achieves in one stage, the ant has high probability of getting the good complete tour. As a notion, this research is performed on three stages. In our investigation, the result of four or more stages should not be reliable.

2.2 Construct Tour

After each ant-bees traveling one stage using the rule of ACS, it backs to hive to interacting with each other (supposing distance from hive to each city=0), and calculating the total amount of pheromone. HABCO also categorizes ant-bees into three types based on pheromone gathered namely scout (S), follower (F), and recruiter (R) as BCO. Scout retains the previous stage and continues the next stage without interacting with the others, follower abandons the previous stage and follows the recruiter, recruiter retains her previous stage and recruits the follower to join her previous stage tour. On HABCO, the judgment of quality of agents is based on the amount pheromone gathered in each stages. Scout agent is firstly chosen 10 % from all agents. The scout agent guarantees the alternative tour is kept on the process. The remainder agents is categorized as follower and recruiter. The probability ant-bee to be follower after traveling one stage is used by Eq. (1).

$$PF = \frac{\tau_{max} - \tau_i}{\sum_{i=1}^{M} (\tau_{max} - \tau_i)}, \quad (1)$$

where i is agent index; M is the total number agent ant-bees; τ_i represents pheromone gathered at each stages by agent; τ_{max} represents pheromone maximum gathered at each stages by agent. Eq. (1) shows that agent which gathers the high pheromone has low probability to be follower. Agent with the highest pheromone (τ_{max}) will absolutely become a recruiter ($PF=0$). After categorized the agents into follower and recruiter, the follower will change her tour and follow one of the follower agent randomly. After exchanging information on hive, the agent ant-bees continue the next stages until finish the tour.

2.3 Pheromone Update

There are 3 types pheromone updating on HABCO. They are local update, semi-global update and global update. Local update is performed as local updating rule in ACS algorithm. After performing one stage, HABCO updates the pheromone

called semi-global update. This update is based on modified ELU-Ants algorithm that ants have more effect on pheromone update where they are in their earlier steps and less effect when they are going to finish the tour.

$$\tau(i,j) = \tau(i,j) + \tau_o . e^{-\frac{5.S}{|S|}}, \quad (2)$$

where S is current stage; $|S|$ is total number stage. Obviously from Eq. (2), the pheromone after performing the first stage ($S=1$) is added more than the next stage. So the ant-bees play fewer impact in update pheromone when they are in their final part of tour. At last, the pheromone on the edges are travelled by the best ant-bee will be updated with global updating. To avoid trapping local optima, the pheromone will be reduced if the best length tour is same in some iterations in a row. In this experiment, the pheromone will be diminished 5% when the best of length tour is same for t times iterations (t=100) in a row. The algorithm of HABCO is described as Fig. 1.

```
Procedure HABCO for solving TSP
     Set parameters and number stage
     For each stages
          For each ant-bees
               Construct Solution and Local update
          End ant-bees
          Semi-global update
          Ranked ant-bees
          For each ant-bees
               //Back to hive
               Classify ant-bees into scout, recruiter, follower
          End ant-bees
          All ant-bees exchange information
          If (stage=number stage)
               Calculating the length complete tour of each ant-bees
               Global update by the best ant-bee
               Ruin local optima
          End if
     End Stage
     Local Search
End
```

Fig. 1. HABCO Algorithm

3 Experiment Results

Experiments are conducted to evaluate the performance of HABCO. The settings of various parameters for the proposed algorithm are described as follows. The number ant bee ($M=10$), number stage ($S=3$), number iteration ($nc=1000$), degree heuristic distance ($\alpha=1$), degree pheromone ($\beta= 2$), pheromone decay ($\gamma=0.1$), and limit exploitation ($q_0 =0.1$).

ACS has an advantage to make the edge more dynamic by perform local update. With local update, pheromone on edge visited is diminished and make it less

desirable. Therefore the premature converge can be avoided. As a consequence, the agent (ant) has many various options to develop the good tour. In other hand, BCO has main advantage on exchange information agents (bees) on their hive. Before backing to hive, construction bee makes the selection candidate node is more static. BCO has also a slightly complex formula than ACS, and also has many variables to set. The difference setting sometimes generates the different significant results. The second comparison should be discussed is comparing HABCO and ACS. Table 1 shows that HABCO outperforms ACS either with 2opt or without 2opt. HABCO also shows slightly better either in average or best case. Although ACS and HABCO uses the same role to construct the tour at each stages, HABCO has an advantage in producing the competitive tour that is generated by their agents. It happens since after agents (ant-bees) finish each stages, HABCO generates the competitive agents by duplicating the good sub tour and discarding the bad sub tour in their hive. As a result on HABCO will have a competitive agents comparing ACS to construct the tour.

Table 1. Performance HABCO ,ACS, and BCO on nine benchmark datasets [15]

Bench-mark	BKS	ACS		ACS+2opt		BCO+2opt (*number bee=10*)		HABCO		HABCO+2opt	
		Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best
Eil51	426	432.4	430	430.3	428	432.1	429	430.7	**428**	**429.1**	**427**
Berlin52	7542	7715.8	**7542**	7611.4	**7542**	7652.6	**7542**	7577.9	**7542**	**7543.5**	**7542**
St70	675	693.4	684	686.2	680	683.2	**675**	686.5	680	**681.1**	**678**
Kroa100	21282	22054.8	21716	21568	21369	21927.3	21369	21789.9	21402	**21413.5**	**21282**
Eil101	629	649.9	641	643.9	636	656.2	645	643	634	**641.7**	**634**
Kroa150	26524	27476.9	26889	27219.1	26912	27452.4	27063	27139.1	26881	**27015.9**	**26781**
D198	15780	16480.2	16202	16027	15918	16351.5	16021	16421.5	16182	**16030**	**15895**
A280	2579	2739.6	2686	2669.9	2640	2834.1	2701	2702.4	2653	**2665.1**	**2630**
U724	41910	46363	44688	44136.4	43781	46516.7	44091	45117.8	43901	**44002.3**	**43432**

4 Conclusion

This paper proposed a new hybrid algorithms called HABCO that is merged algorithms ACS, BCO and ELU-Ants. HABCO splits the tour into some stages likes on BCO algorithm. However this proposed method mimics ACS to construct the tour and perform local and global updating. The quality of each agents HABCO is judged based on the pheromone gathered at each stages. The pheromone gathered classifies agent into follower, recruiter and scout. A good judgment on premature step makes the probably good agent (identified as recruiter) duplicated, and the probability bad agent (identified as follower) is discarded. The scout agent is exist to maintain the new alternative tour. This process produce the competitive agent in each iteration. This algorithm also proposed the semi-global update pheromone to adopt the ELU-Ants algorithm. Our simulation results shows that the proposed algorithm, HABCO, outperforms ACS and BCO on several data sets, either with or without local search 2opt.

References

1. Dorigo, M., Maniezzo, V., Colorni, A.: The ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics Part B 26(1), 2941 (1996)
2. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation 1(1), 5366 (1997)
3. Stutzle, T., Hoos, H.H.: Improving the Ant System: A Detail Report on the MAXMIN Ant System. Technical Report. AIDA-96-12. FG Intellektik, FB Informatik, TU Darmstadt, Germany (1996)
4. Naimi, H.M., Taherinejad, N.: New robust and efficient ant colony algorithms: Using new interpretation of local updating process. Expert Systems with Applications 36(1), 481–488 (2009)
5. Chen, S.M., Chien, C.Y.: Parallelized genetic ant colony systems for solving the traveling salesman problem. Expert Systems with Applications 38(4), 3873–3883 (2011)
6. Chen, S.M., Chien, C.Y.: Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques. Expert Systems with Applications 38(12), 14439–14450 (2011)
7. Sjoerd, V.D.Z., Marques, C.: Ant colony optimization for job shop scheduling. In: Proceedings of Workshop on Genetic Algorithms and Artificial Life GAAL (1999)
8. Gambardella, L.M., Taillard, E.D., Dorigo, M.: Ant colonies for the quadratic assignment problem. Journal of the Operational Research Society 50(2), 167–176 (1999)
9. Lucic, P.: Modeling transportation problems using concepts of swarm intelligence and soft computing. PhD Thesis Civil Engineering Virginia Polytechnic Institute and State University (2002)
10. Teodorovic, D., Lucic, P., Markovic, P., Orco, M.D.: Bee colony optimization: principles and applications. In: 8th Seminar on Neural Network Applications in Electrical Engineering, NEUREL (2006)
11. Lucic, P., Teodorovic, D.: Vehicle routing problem with uncertain demand at nodes: the bee system and fuzzy logic approach. In: Verdegay, J.-L. (ed.) Fuzzy Sets Based Heuristics for Optimization. STUDFUZZ, vol. 126, pp. 67–82. Springer, Heidelberg (2003)
12. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of Global Optimization 39(3), 459–471 (2007)
13. Benatchba, K., Admane, L., Koudil, M.: Using bees to solve a data-mining problem expressed as a max-sat one. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2005. LNCS, vol. 3562, pp. 212–220. Springer, Heidelberg (2005)
14. Chong, C.S., Low, M.Y.H., Sivakumar, A.I., Gay, K.L.: A bee colony optimization algorithm to job shop scheduling. In: Proceedings of Winter Simulation Conference, pp. 1954–1961 (2006)
15. TSPLIB (2012), http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/tsp

Author Index

Ahn, Sang-Ho 541, 549
Akhmedjanov, Umid 347
An, Xin 253
Atkinson, John 421

Bang, Hyo-Chan 87, 125, 457, 463
Bng, Hyo Jung 451

Chang, Jae-woo 293, 301
Chang, Jed Kao-Tung 557
Chen, Jiancheng 253
Chen, Sheng-Chang 509
Chen, Shih-Yue 509
Chiang, Ming-Chao 615, 621, 629
Cho, Chang-Woo 335
Cho, Dong-Sub 25
Cho, Gi-hwan 433
Cho, Ilkwon 69
Cho, Shung Han 165
Cho, Yongwon 371
Cho, Young Sung 441
Choi, Dong-hoon 293, 301, 307
Choi, Jae-Yeong 117
Choi, Jinsung 391
Choi, Junyoung 287
Choi, Ki-Young 335
Choi, Maengsik 225
Choi, Okkyung 353, 365, 391
Choi, Seokjin 535
Choi, Seowon 99
Choi, Sung-Pil 239, 267, 273
Chuang, Po-Jen 637
Chung, In-Jeong 143
Chung, Tai-Myoung 111
Chung, Yeh-Ching 577
Chung, Yeonwoo 105
Chung, Yongwha 105, 157
Chung, Youngseok 535
Cui, Yun 353, 359, 391

Degefa, Fikadu B. 521

Englmeier, Kurt 421

Gil, Joon-Min 31
Gim, Jangwon 191, 261
Girsang, Abba Suganda 643
Gu, Yi 359
Guo, Jianfeng 231
Guu, Tyng-Tyng 81

Haghighi, Mo 173
Han, Byong-John 329
Han, Seung Ho 359, 365
Havinga, Paul 397
Hong, Sangjin 165
Hong, Seungtae 301
Hong, Sugwon 307
Hosen, A.S.M. Sanwar 433
Hsiao, Yuan-Kai 569
Hsieh, Meng-Yen 93, 563
Hsieh, Ting-An 609
Hsu, Ching-Hsien 609
Hsu, Kuo-Hsun 609
Huang, Bo-Chi 621
Huang, Kuo-Chan 609
Huang, Wei-Cheng 629
Hwang, Myunggwon 191, 261
Hwang, Woomin 315

Ihm, Sun-Young 585
In, Kwanho 39
Islam, Md Shohidul 47, 591

Jang, Illwoong 371
Jeon, Dongwoon 341
Jeong, Chang-Hoo 267
Jeong, Chang-Sung 329, 335
Jeong, Do-Heon 191, 261
Jeong, Gowun 475
Jeong, Young-Sik 585
Jiang, Jiulei 501
Jin, Qun 501
Jo, Heeseung 293, 301
Jo, Hoon 75
Joo, Bok-Gyu 143
Jung, Euihyun 69
Jung, Hanmin 191, 239, 247, 261, 267, 273, 287
Jung, HaRim 31
Jung, In-Yong 329
Jung, Jong-jin 353, 365
Jung, Sung-Jae 247
Jung, Sung-Min 111

Kang, Ana 99
Kang, Daehyun 143
Kang, Myeongsu 599
Kang, Sun Moo 69
Kang, Taegyeong 463
Kang, Taekyeong 87, 125
Kim, Byung-Gyu 117
Kim, Cheol-Hong 599
Kim, Daesung 371
Kim, Doo-Hyun 341
Kim, Eunhye 279
Kim, Eun Yi 347
Kim, Haelyeon 105
Kim, Harksoo 219, 225
Kim, Heegon 157
Kim, Hyeong-il 293
Kim, Ikkyun 489
Kim, Jeongyeun 515, 529
Kim, Jinhyung 191, 261
Kim, Jiwon 13
Kim, Jong-Ho 541, 549
Kim, Jong-Myon 47, 591, 599
Kim, Kangseok 391
Kim, Keon Uk 99
Kim, Ki-Hyun 335
Kim, Mijin 483
Kim, Moonhyun 469
Kim, Myoungjin 353, 359, 365, 391
Kim, Nam-Uk 111
Kim, Sang-Kyoon 541, 549
Kim, Sehun 279
Kim, Seongkyu 31, 39
Kim, Sung-Ki 117
Kim, Ung-Mo 31, 39
Kim, Woongsup 377
Kim, Yangwoo 377
Kim, Yilip 515, 529
Kim, Youngsoo 489
Ko, Eunjeong 347
Ko, Gunhwan 321
Kum, Seung-woo 353
Kwon, Kyunglag 143

Lai, Kuan-Chou 577, 609
Lee, Daesung 287
Lee, Dong-Young 111
Lee, Gangin 19
Lee, Ha-Kyung Jennifer 25
Lee, Hanku 353, 359, 365, 391
Lee, Hyungkyu 87, 125, 457, 463, 515
Lee, Hyun-Jung 495
Lee, Myungho 307
Lee, Myung-Joon 55
Lee, Seungha 377
Lee, Seung-Hyun 111
Lee, Seungwoo 239, 247, 267
Lee, Sungju 105, 157
Lee, Youngsook 495, 535
Lee, Yunjin 385
Li, Kuan-Ching 93, 563, 577, 609
Li, Weifeng 211
Li, Weimin 501
Li, Yi-Hsung 509
Lim, Mingyu 385
Lim, Si-Hyung 451
Lim, Tae-Beom 353
Lin, Pei-Jung 509
Lin, Yen-Wen 569
Liu, An-Peng 81
Liu, Chen 557
Liu, Dongpei 413
Liu, Hengzhu 413
Lu, Ssu-Hsuan 577

Mazhar, Sajjad 191
Min, Byoung-Joon 117
Min, Jun-Ki 151
Moon, Nammee 165
Moon, Song Chul 441
Moon, Sujung 371
Mothe, Josiane 421
Murtagh, Fionn 421

Nasridinov, Aziz 585

Oh, Jeong Seok 451
Oh, Wonyeong 99

Park, Daihee 105
Park, Dong-Hwan 87, 125
Park, Jong-Eun 55
Park, Jongmoon 55
Park, Jun-Jae 549
Park, Ki-Woong 315
Park, Kyongseok 321
Park, Kyu Ho 315
Park, Namje 87, 125, 457, 463, 483, 489, 515, 529
Park, Soon-cheol 75
Park, Young-Ho 585
Park, Young-Su 541
Pereira, Javier 421
Pyeon, Muwook 371
Pyun, Gwangbum 1

Qi, Jiandong 197
Qiao, Xiaodong 239

Ryang, Heungmo 7

Scholten, Hans 397
Seo, Dong-min 247
Seo, SeungHyun 359
Seo, Yeon-gyeong 99
Seo, Yong-Ho 475
Shi, Chen 231
Shi, Qingwei 239
Shieh, Wann-Yun 81
Shin, Junsoo 219, 225
Shin, Sungho 273
Shin, Youngsung 301
Shin, Yunhee 347
Shoaib, Muhammad 179
Sohn, Jongsoo 143
Son, Jeonghan 99
Song, Sa-kwang 191, 261
Song, Wang-Cheol 179
Song, Yeongkil 219
Song, Yong 315
Sug, Hyontai 185

Tran, Nhat-Phuong 307
Tsai, Chun-Wei 621, 629, 643
Tsai, Yin-Te 93
Tseng, Shih-Pang 615
Turkes, Okan 397

Um, Jung-Ho 267, 273

Vladimir, Tyan 341

Wang, Xiaodong 405
Wang, Zehan 197
Won, Dongho 483, 495, 521, 535
Won, Jongjin 469
Wu, Meng-Syue 509

Xia, Xiao 405
Xu, Bingqing 63, 131, 137
Xu, Shuo 191, 211, 239, 261, 273

Yang, Chu-Sing 615, 637, 643
Yang, Hyun S. 475
Yeh, Ching-Hung 93, 563
Yeh, Hongjin 391
Yeo, Sang-Soo 475
Yoon, Jongcheol 321
Yoon, Kee-Hyung 25
Yoon, Min 293
Yun, Unil 1, 7, 13, 19
Yun, Young-Mi 25

Zhang, Haodong 211
Zhang, Lichen 63, 131, 137
Zhang, Yuan 253
Zhang, Yunliang 205
Zhao, Xiaohua 501
Zhou, Li 413
Zhou, Xiaokang 501
Zhou, Xingming 405
Zhu, Lijun 191, 197, 211, 231, 239, 261, 273
Zhu, Tao 405